12144

MATHEMATICAL APPLICATIONS OF TEST OF HYPOTHESIS AND SAMPLING TECHNIQUE

K.M. MANIKANDAN

E
515
M 314

MATHEMATICAL APPLICATIONS OF TEST OF HYPOTHESIS AND SAMPLING TECHNIQUE

K.M. MANIKANDAN Msc., M.phil., PGDCA.

Senior Lecturer

Department of Mathematics with Computer Applications

Dr. SNS Rajalakshmi College of Arts & Sciences

Chinnavedampatti, Coimbatore - 641 006

Published by

VARUN PUBLICATIONS

Salem.

MATHEMATICAL APPLICATIONS OF TEST OF HYPOTHESIS AND SAMPLING TECHNIQUE

125/-

E
515
M314

Author	:	K.M. Manikandan
Language		English
Edition		First Edition, July 2008
Size		Dummy 1/8
Point		11
Copy Right ©		K.M. Manikandan
ISBN No		**978-81-907472-3-3**
Pages	:	144 + V
Copies	:	1000
Price	:	**125**/-
Wrapper & Typesetting	:	Rainbow Publications
Paper	:	Maplito
Printer	:	Rainbow publications, CBE - 43
Publishers	:	**Varun Publications** Ponnagar, Metturdam, Salem - 634 601 E-mail : varunindu@yahoo.co.in

12144
DBA000109406ENG

PREFACE

This book entitled **Mathematical Applications of test of Hypothesis and Sampling Technique** is mainly written to cover the areas of test of Hypothesis and sampling technique.

A conscious effort has been made to make the presentation of all topics very simple and to develop the subject - matter in step - by - step approach, such that the book can serve as self study text for the students

This book will be useful to the students of Mathematics and Statistics, who study their subjects in ancillary and major discipline.

I hope that comprehensive treatment of the topics covered in this book will be welcomed both by the teachers and students.

Suggestions for the improvement of this book is most welcome and will be acknowledged most gratefully

- Author

Contents

Chapter I

Test Of Hypothesis For Large And Small Sample

1.1 Test Of Hypothesis 1

1.2. Power Of The Test 6

1.3 Most Powerful Test 9

1.4 Uniformly Most Powerful Test 11

1.5 Neyman - Pearson Lemma 12

1.6 Unbiased Test And Unbiased Critical Region 15

1.7 Optimum Regions Ans Sufficient Statistics 18

1.8 Likelihood Ratio Test 18

1.9 Properties Of Likelihood Ratio Test 23

1.10 Test Of Significance For Large Samples 24

1.11 Test Of Significance For Single Mean 28

1.12 Test Of Significance For Difference Of Means 34

1.13 Test Of Significance For Single Proportion 42

1.14 Test Of Significance For Difference Between Two Sample Proportions 47

1.15 Test Of Significance For Difference Between Two Sample Standard Deviations 55

Exercise 57

Chapter II

2.1 Test of significance for small samples - t Test for small sample 59

2.2 Paired 't' Test For Single Mean 66

2.3 Significance Of Difference between Two Small Sample Means 70

2.4 Snedceor's F-Test of Significance(For Small Samples) 78

2.5 χ^2 Test 85

2.6 χ^2 Test of Independence of attributes 95

Exercise 103

Chapter III

Sampling Technique

3.1 Introduction 108

3.2 Simple Random Sampling 111

3.3 Systematic Sampling 113

3.4 Stratified Random Sampling 113

3.5 Standard Error 114

3.6 Sampling with and without Replacement 116

3.7 Theorems on simple random sampling 124

3.8 SRSWOR VS SRSWR 132

3.9 Sampling Errors 133

3.10 Non Sampling Errors 135

CHAPTER I

TEST OF HYPOTHESIS FOR LARGE SAMPLES

1.1 TEST OF HYPOTHESIS

The main problems in statistical inference can be broadly classified into two areas.

i. The area of estimation of population parameters and setting up of confidence intervals for them, ie the area of point and interval estimation and

ii. Test of statistical hypothesis

Statistical Hypothesis : Simple and composite

Definition :

A statistical hypothesis is some statement (or) ascertain about a population or equivalently about the population distribution characterizing a population which we want to verify on the information available from the sample. If the statistical hypothesis specifies the population completely then it is termed as **statistical hypothesis**, otherwise it is called a **composite statistical hypothesis**.

Example :

Let $X_1, X_2 \ldots\ldots X_n$ be a random sample of size 'n' from a normal population with mean μ and variance σ^2 then, the hypothesis $H_0 : \mu = \mu_0 , \sigma^2 = \sigma_0^2$ is a simple hypothesis, whereas each of the following is a composite hypothesis.

i. $\mu = \mu_0$ ii. $\sigma^2 = \sigma_0^2$ iii. $\mu < \mu_0 , \sigma^2 = \sigma_0^2$

iv. $\mu > \mu_0, \sigma^2 = \sigma_0^2$ v. $\mu = \mu_0, \sigma^2 < \sigma_0^2$

vi. $\mu = \mu_0, \sigma^2 > \sigma_0^2$ vi. $\mu < \mu_0, \sigma^2 > \sigma_0^2$

A hypothesis which does not specify completely 'r' parameters of a population is termed as a **composite hypothesis** with r- degrees of freedom.

Test of a statistical hypothesis

A test of a statistical hypothesis is a two action decision problem after the experimental sample values have been obtained, the two – actions being acceptance (or) rejection of hypothesis under consideration.

Critical Region:

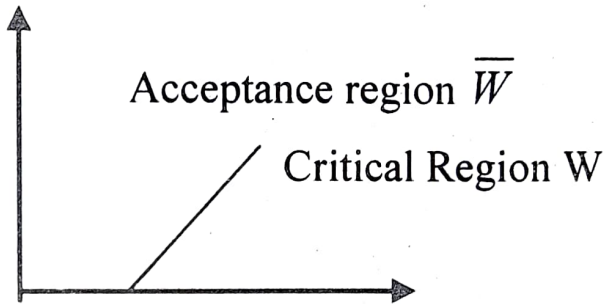
Let $X_1, X_2....X_n$ be the sample observations denoted by O. All the values of O will be aggregate of a sample and they constitute a space called the **sample space** which is denoted by S.

Since the sample values $X_1, X_2.....X_n$ can be taken as a point in n- dimensional space, we specify some region of the n-dimensional space and see whether this point lies within this region or outside this region. We divide the whole sample space S into two disjoint parts. W and S-W or $\overline{W}$ and W' The null hypothesis H_0 is rejected if the observed sample point falls in W and if it falls in $\overline{W}$ we reject H_1 and accept H_0.

The region of rejection of H_0, when H_0 is true is that the region of the outcome set where H_0 is rejected if the sample

point falls in that region and is called **critical region.** The size of the critical region is α.

Suppose if the test is based on a sample of size 2, then the outcome set (or) the sample space is the first quadrant in a 2-dimensional space and a test criterion will enable us to separate our outcome set into two complementary subsets W and $\overline{W}$. If the sample point falls in the subset W, H_0 is rejected, otherwise H_0 is accepted.



Errors of type I and type II

The error of rejecting H_0 (accepting H_1) when H_0 is true is called **type I error** and the error of accepting H_0 when H_0 is false (H_1 is true) is called **type II error**. The probabilities of type I and type II errors are denoted by α and β respectively.

α = probability of type I error

= probability of rejecting H_0 when H_0 is true

β = probability of type II error

= probability of accepting H_0 when H_0 is false

Symbolically

$P(x \varepsilon W | H_0) = \alpha$ where $X = (X_1, X_2 \ldots\ldots X_n)$

$$\Rightarrow \int_{W} L_o dx = \alpha$$

where L_o is the likelihood function of the sample observations under H_0 and

$\int dx$ represent the n- fold integral

$$\int \int \ldots\ldots \int dx_1, dx_2 \ldots\ldots dx_n$$

Again $P(x \varepsilon \overline{W} | H_1) = \beta$

$$\Rightarrow \int_{\overline{w}} L_1 dx = \beta$$

where L_1 is the likelihood function of the sample observations under H_1.

Since $\int_{W} L_1 dx + \int_{\overline{w}} L_1 dx = 1$

$$\int_{W} L_1 dx = 1 - \int_{\overline{w}} L_1 dx$$

$$\int_{W} L_1 dx = l - \beta$$

$\Rightarrow P(x \in W| = 1- \beta$

Level of significance : α the probability of type I error is known as the level of significance of the test. It is also called the size of the critical region.

Example 1.1.1

If $x \geq 1$, is the critical region for testing $H_0 : \theta = 2$

against the alternative $\theta = 1$, on the basis of single observation from the population

$f(x,\theta) = \theta \exp(-\theta x),\ 0 \le x \le \infty$.

Obtain the values of type I and type II errors.

Solution :

Here $W = \{x : x \ge 1\}$ $\overline{W} = \{x : x < 1\}$ and $H_0 : \theta = 2,\ H_1 : \theta = 1$

α = size of type I error

$= P(x \in W \mid H_0)$

$= P(x \ge 1 \mid \theta = 2)$

$$= \int_1^{\infty} [f(x,\theta)]_{\theta=2}\, dx$$

$$= \int_1^{\infty} 2e^{-2x} dx$$

$$= 2.\left(\frac{e^{-2x}}{-2}\right)^{\infty} = 1/e^2$$

β = size of type II error

$= P(x \in \varepsilon\ \overline{W} \mid H_1)$

$= P(x < 1 \mid \theta = 1)$

$$= \int_0^1 e^{-x} dx = -(e^{-x})_0^1$$

$$= -(e^{-1} - e^{-0})$$

$$= 1 - e^{-1} = \frac{e-1}{e}$$

1.2 POWER OF THE TEST

$1-\beta$ defined in $P(x \in W | H_1) = 1 - \beta$ is called the power function of the test hypothesis H_0 against the alternative hypothesis H_1 The value of the power function at a parameter point is called the **power of the test** at that point.

Note : 1

In quality control terminology, α and β are termed as producer's risk and consumer's risk respectively.

Example 1.2.1

Let X have a p.d.f of the form $f(x,\theta) = \frac{1}{\theta} e^{-x/\theta}$, $0 < x < \infty$, $\theta > 0$

$= 0$, else where

To test $H_0 : \theta = 2$ against $H_1 : \theta = 1$, use random sample x_1, x_2 of size 2 and define a critical region $W=\{(x_1, x_2): 9.5 \leq x_1 + x_2\}$ Find (i) power of the test (ii) significance level of the test.

Solution : We are given the critical region

$$W = \{(x_1,x_2) : 9.5 \leq x_1+x_2\}$$

$$= \{(x_1, x_2): x_1 + x_2 \geq 9.5)\}$$

Size of the critical region (or) the significance level of the test is given by

$$\alpha = P(x \in W \mid H_0)$$

$$= P(x_1 + x_2 \geq 9.5 \mid H_0)$$

In sampling from the given exponential distribution

$$\frac{2}{\theta} \sum_{i=1}^{n} x_i \sim \chi^2(2n)$$

$\Rightarrow U = \frac{2}{\theta}(x_1 + x_2) \sim \chi^2(4)\ (n = 2)$

Therefore $\alpha = P\left[\frac{2}{\theta}(x_1 + x_2) \geq \frac{2}{\theta}(9.5) \mid H_o\right]$

$= P[\chi^2_{(4)} \geq 9.5]$ [Q therefore under $H_0 : \theta = 2$]

$\Rightarrow \alpha = 0.05$ (from probability tables of χ^2 distribution)

Power of the test is given by

$1 - \beta = P(x \in W \mid H_1)$

$= P(x_1 + x_2 \geq 9.5 \mid H_1)$

$= P\left(\frac{2}{\theta}(x_1 + x_2) \geq \frac{2}{\theta}9.5 \mid H_1\right)$

$= P(\chi^2(4) \geq \frac{2}{1}9.5)$ (Q under $H_1 : \theta = 1$)

$= P(\chi^2(4) \geq 19)$

Example 1.2.2

Let p be the probability that a coin will fall head in a single toss in order to test $H_0 : p = \frac{1}{2}$ against $H_1 : p = \frac{3}{4}$. The coin is tossed 5 times and H_0 is rejected if more than 3 heads are obtained. Find the probability of type I error and power of the test.

Solution :

$H_0 : p = \frac{1}{2}$ and $H_1 : p = \frac{3}{4}$ If r.v x denotes the no of heads in 'n'

tosses of a coin then X ~ B (n,p) so that

$$P(X = x) = {}^{n}C_{x} \quad p^{x} q^{n-x.} = {}^{5}C_{x} \; p^{x} q^{n-x}$$

The critical region is given by

$W = \{ x: x \geq 4\}$

$\Rightarrow \overline{W} = \{x : x \leq 3\}$

α = probability of type I error

$= P (x \geq 4 \mid H_0)$

$$= P\left(x = 4 \mid p = \frac{1}{2}\right) + P\left(x=5 \mid p = \frac{1}{2}\right)$$

$$= {}^{5}C_{4}\left(\frac{1}{2}\right)^{4}\left(\frac{1}{2}\right) + {}^{5}C_{5}\left(\frac{1}{2}\right)^{5}$$

$$= 6\left(\frac{1}{2}\right)^{5} = \frac{3}{16}$$

β = Probablity of type II error

$= P (x \in \overline{W} \mid H_1)$

$= 1 - P (x \in W \mid H_1)$

$$= 1 - P\left(x = 4 \mid p = \frac{3}{4}\right) + P(x = 5 \mid p = \frac{3}{4})]$$

$$= 1- \left[{}^{5}C_{4}\left(\frac{3}{4}\right)^{4}\left(\frac{1}{4}\right) + {}^{5}C_{5}\left(\frac{3}{4}\right)^{5} \right]$$

$$= 1 - \left(\frac{3}{4}\right)^{4}\left[\frac{5}{4} + \frac{3}{4}\right]$$

$$= 1 - \frac{81}{128} = \frac{47}{128}$$

Therefore power of the test $1-\beta = \frac{47}{128}$

1.3 MOST POWERFUL TEST (MP TEST)

Let us consider the problem of testing a simple hypothesis $H_0 : \theta = \theta_0$ against simple alternative hypothesis

$H_1 : \theta = \theta_1$.

Definition :

The critical region W is the Most Powerful (MP) critical region of size α (and the corresponding test a most powerful test of level α) for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ if

$$P(x \in W|H_1) = \int_W L_0 dx = \alpha \qquad \ldots (1)$$

and $P(x \in W|H_1) \geq P(x \in W_1|H_1)$ for every other critical region satisfying equation (1)

Example 1.3.1

Given he frequency function $f(x, \theta) = \frac{1}{\theta}$, $0 \leq x \leq 0$

$= 0$, else where

and test the null hypothesis $H_0 : \theta = 1$ against $H_1 : \theta = 2$ by means of single observed value of x. Find the size of type I and type II errors by choosing the interval (i) $0.5 \leq x$ (ii) $1 \leq x \leq 1.5$ as the critical regions? Also obtain power function of the test

Solution :

Here we want to test,

$H_0 : \theta = 1$ against $H_1 : \theta = 2$

(i) Here $W = (x: 0.5 \leq x\}$

$= \{ x: x \geq 0.5\}$

$\overline{W} = \{x: x \leq 0.5\}$

$\alpha = P (x \in W \mid H_0\}$

$= P (x \geq 0.5 \mid \theta = 1\}$

$= P (0.5 \leq x \leq \theta \mid \theta = 1\}$

$= P(0.5 \leq x \leq 1 \mid \theta = 1\}$

$$= \int_{0.5}^{1} [f(x,\theta)]_{\theta=1} dx$$

Similarly

$\beta = P(x \in \overline{W} \mid H_1)$

$= P (x \leq 0.5 \mid \theta = 2\}$

$$= \int_{0}^{0.5} [f(x,\theta)]_{\theta=2} dx$$

$$= \int_{0}^{0.5} \frac{1}{2} dx = 0.25$$

The sizes of type I and type II errors are respectively $\alpha = 0.5$, and $\beta = 0.25$ and power function of the rest = $1- \beta = 0.75$

i. $W = \{x: 1 \leq x \leq 1.5\}$

$\alpha = P(x \in W \mid \theta = 1\}$

$$= \int_{1}^{1.5} [f(x,\theta)_{\theta=1} dx$$

$= 0$ ($\because$ under $H_o : \theta = 1$, $f(x,\theta) = 0$, for $1 \leq x \leq 1.5$]

$$\beta = P(x \in \bar{w} | \theta = 2)$$

$$= 1 - P(x \in W | \theta = 2)$$

$$= 1 - \int_{1}^{1.5} [f(x,\theta)]_{\theta=2} dx$$

$$= 1 = \int_{1}^{1.5} \frac{1}{2} dx$$

$= 0.75$

Therefore function $1 - \beta = 1 - 0.75 = 025$.

1.4 UNIFORMLY MOST POWERFUL TEST (UMP TEST)

Let us now consider the case of testing a simple null hypothesis against a composite alternative hypothesis. eg. of testing $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \neq \theta_0$. In such a case, for a predetermined α the best test for H_0 is called the uniformly most powerful test of level α.

Definition :

The region W is called **uniformly most powerful** (UMP) critical region of size α . (and the corresponding test a uniformly most powerful test of α) for testing as uniformly most powerful test of α) for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$

(i.e) $H_1 : \theta \neq \theta_0$ if $p(x \varepsilon W | H_0) = \int_W L_o dx = \alpha$ ...(2)

and $p(x \in W|H_1) \geq P(x \in W_1|H_1)$ for all $\theta_1 \neq \theta_0$ whenever the region satisfies equation (2)

1.5 NEYMAN -PEARSON LEMMA

Let $k > 0$, be a constant and W be a critical region of size α such that

$$W = \{ x \varepsilon S : \frac{f(x,\theta_1)}{f(x,\theta_0)} > k\}$$

$$\Rightarrow W = \{ x \varepsilon S : \frac{L_1}{L_0} > k\} \quad \text{and} \quad \overline{W} = \{ x \varepsilon S : \frac{L_1}{L_0} \leq k\}$$

Where L_o and L_1 are the likelihood functions of the sample observations $x = (x_1, x_2 \ldots\ldots\ldots x_n)$ under H_0 and H_1 respectively. Then W is the most powerful critical region of the test hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \theta_1$

Proof :

we are given $P(x \in W | H_0) = \int_W L_o dx = \alpha \qquad \ldots(1)$

The power of the region is $P(x \in W | H_1) = \int_W L_1 dx = 1 - \beta \quad \ldots.(2)$

In order to establish the lemma, we have to prove that there exists no other critical region of size less than or equal to α, which is more powerful than W.

Let W_1 be another critical region of size $\alpha_1 \leq \alpha$, and power $1 - \beta_1$ that we have

$$p(x \in W_1 | H_0) = \int_{W_1} L_0 dx = \alpha_1 \quad \ldots.(3)$$

and $p(x \in W_1 \mid H_1) = \int_{W_1} L_1 dx = 1 - \beta_1$ ....(4)

Now have to prove that $1 - \beta \geq 1 - \beta_1$

Let W = AUC and W_1 = BUC

(Diagram)

(c may be empty, i.e, W and W_1 may be disjoint)

If $\alpha_1 \leq \alpha$. We have

$$\int_{W_1} L_0 dx \leq \int_{W} L_0 dx$$

$$\Rightarrow \int_{BUC} L_0 dx \leq \int_{AUC} L_0 dx$$

$$\Rightarrow \int_{B} L_0 dx \leq \int_{A} L_0 dx$$

$$\Rightarrow \int_{A} L_0 dx \geq \int_{B} L_0 dx. \qquad \text{.....(5)}$$

Since A C W,

$$\Rightarrow \frac{\int_{A} L_1 dx}{\int_{A} L_0 dx} > k$$

$$\Rightarrow \int_{A} L_1 dx > k \int_{A} L_0 dx$$

$$\geq k \int_{B} L_o dx. \qquad \text{.....(6)} \qquad (\because \text{ from (5) })$$

Also from $\frac{L_1}{L_0} \leq k \;\; x \in \overline{W}$

$$\Rightarrow \int_{\overline{W}} L_1 dx \le k \int_{\overline{W}} L_0 dx$$

$$[\because From 6K \int_B L_0 dx \le k \int_A L_1 dx]$$

The result also holds for nay subset of $\overline{W}$, say $\overline{W} \cap W_1 = B$.

Hence $\int_B L_1 dx \le k \int_B L_0 dx$

$$\int_B L_1 dx \le k \int_A L_1 dx$$

Adding $\int_c L_1 dx$ to both sides

$$\int_B L_1 dx \le \int_C L_1 dx \le \int_A L_1 dx \le \int_C L_1 dx$$

$$\int_{B \cup C} L_1 dx \le \int_{AUC} L_1 dx$$

$$\int_{W_1} L_1 dx \le \int_W L_1 dx$$

$1-\beta_1 \le 1-\beta$

(i.e) $1-\beta \ge 1-\beta_1$

Hence the lemma

1.6 UNBIASED TEST AND UNBIASED CRITICAL REGION

Let us consider the testing $H_0 : \theta = \theta_0$ against $H_1: \theta = \theta_1$. The critical region W and consequently the test based on it is said to be unbiased if the power of the test exceeds the size of the critical region (ie.) if power of the test $\geq$ size of CR

$\Rightarrow 1-\beta \geq \alpha$

$\Rightarrow P_{\theta_1}(W) \geq P_{\theta_0}(W)$

$\Rightarrow P[x:x \in W|H_1) \geq P[x:x \in W|H_0]$

In other words, the critical region W is said to be unbiased if

$P_{\theta_0}(W) \geq P_{\theta_0}(W) \quad \forall\ \theta\ (\neq \theta_0) \in \Theta$

Theorem : 1.6.(i)

i. Every most powerful (MP) or uniformly most powerful (UMP) critical region is necessarily unbiased.

i. If W be a MPCR of size α for testing $H_0 : \theta = \theta_o$ against $H_1 : \theta = \theta_1$ then it is necessarily unbiased.

ii. Similarly if W be UMPCR of size α for testing $H_0 : \theta = \theta_o$ against $H_1 : \theta \in \in \Theta$, then it also unbiased.

Proof :

Since W is an MPCR of size α for testing $H_0 : \theta = \theta_o$ against $H_1 : \theta = \theta_1$ by Neymann Pearson's Lemna, we have $\forall\ k > 0$,

$$W = \{x : L(x, \theta_1) \geq kL(x,\theta_0)\}$$

$$= \{x : L_1 \geq k\, L_o\}$$

and $W' = \{x: L(x, \theta_1) < k\, L(x, \theta_0)\}$

$= \{x: L_1 < kL_0\}$ where k is determined so that the size of the test is α.

$P_{\theta_0}(W) = P(x \in W \mid H_0)$

$$= \int_W L_o dx = \alpha \quad \ldots\ldots\ldots\ldots\ldots\ldots\text{(i)}$$

To prove that W is unbiased, we have to show that power of $W \geq \alpha$

$P_{\theta 1}(W) \geq \alpha.$

We have $P_{\theta 1}(W) = \int_W L_1 dx \geq k \int_W L_o dx = k\alpha.$ [$\because$ on W, $L_1 \geq kL_0$]

(i.e.) $P_{\theta 1}(W) \geq k\alpha \quad \forall \kappa > 0$

Also, $1 - P_{\theta 1}(W) = 1 - P(x \in W \mid H_1)$

$= P(x \in W' \mid H_1)$

$$= \int_{W'} L_1 dx$$

$$1 - P_{\theta 1}(W) = \int_{W'} L_1 dx$$

$$< k \int_{W'} L_o dx \quad (\because \text{ on } W', L_1 < k L_0)$$

$= k\,P(x: x \in w' / H_o)$

$= k[1 - P(x: x \in W \mid H_o)]$

$= k[1 - \alpha]$

(i.e) $1 - P_{\vartheta}(W) < k(1-\alpha), \quad \forall k > 0$ ..... (iii)

Case (i)

Assume $k \geq 1$ then from (ii) we get, $P_{\theta_1}(W) \geq k$

$\alpha \geq \alpha$

$\Rightarrow$ W is unbiased CR.

Case (ii)

$0 < k < 1$. If $0 < k < 1$, then from

(iii) we get $1 - P_{\theta_1}(W) < 1 - \alpha$

$\Rightarrow P_{\theta_1}(W) > \alpha$

$\Rightarrow$ W is unbiased CR.

Therefore MP critical region is unbaised.

(ii) If W is UMPCR of size α, then also proof holds if for θ_1 we write θ such that $\theta \in \Theta$ so we have $P_\theta(W) > \alpha \quad \forall \ \theta \in \Theta$

Hence W is unbiased CR.

1.7 OPTIMUM REGIONS AND SUFFICIENT STATISTICS

Let $x_1, x_2 .. x_n$ be a random sample of size 'n' from a population with p.d.f $f(x,\theta)$ where the parameter θ may be a vector. Let T be sufficient statistic for θ. Then by factorization theorem,

$$L(x,\theta) = \prod_{i=1}^{n} f(x_i,\theta)$$

$$= g_\theta (t(x)).\ h(x) \ldots\ldots\ldots\ldots (i)$$

where $g_\theta (t(x))$ is the marginal distribution of the statistic $T = t(x)$.

By Neymann pearson's lemma, the MPCR for testing $H_0 : \theta = \theta_0$ against $H_1: \theta = \theta_1$ is given by

$$W = \{x : L(x, \theta_1) \geq k\ L(x, \theta_0)\} \quad \forall\ k > 0 \ \ldots\ldots\ldots (ii)$$

From (i) and (ii)

$$W = \{x : g_{\theta_1}(t(x)).\ h(x) \geq k\ g_{\theta_0}(t(x)).\ h(x))\} \quad \forall\ k > 0$$

$$= \{ x: g_\theta (t(x)) \geq k\ g_{\theta_0}(t(x)).\ h(x))\} \quad \forall\ K > 0$$

Hence if $T = t(x)$ is sufficient statistic for θ then MPCR for the test may be defined in terms of the marginal distribution of $T = t(x)$.

1.8 LIKELIHOOD RATIO TEST

Neymann – Pearson Lemma based on the magnitude of the ratio of two probability density functions provides the best test for testing simple hypothesis against simple alternative hypothesis. The best test in any given situation depends on the nature of the

population distribution and the form of the alternative hypothesis being considered. Likelihood ratio test was introduced by Neymann and Pearson for testing a hypothesis, simple or composite, against a simple or composite alternative hypothesis. This test is related to the maximum likelihood estimates.

Parameter space :

Let us consider a random variable X with p.d.f $f(x,\theta)$. In most common applications, though not always, the functional form of the population distribution is assumed to be known except for the value of some unknown parameter (s) θ, which may take any value on a set Θ. This is expressed by writing the p.d.f in the form $f(x, \theta)$, $\theta \in \Theta$. The set which is the set of all possible values of θ is called the parameter space. Such a situation gives rise not one probability distribution but a family of probability distributions which we write as $\{f(x, \theta), \theta \in \Theta\}$

For example if $X \sim N(\mu,\sigma^2)$ then the parameter space

$$\Theta = \{(\mu,\sigma^2) : -\infty < \mu < \infty, 0 < \sigma < \infty\}$$

In particular for $\sigma^2 = 1$, the family of probability distributions is given by

$$\{N(\mu,1): \mu \in \Theta\} \text{ where } \Theta = \{\mu: -\infty < \mu < \infty\}$$

Consider a general family of distributions

$$\{f(x : \theta_1, \theta_2 \ldots\ldots\ldots\ldots\ldots\ldots \theta_k) : \theta_i \in \Theta, \ i = 1,2,\ldots\ldots\ldots\ldots\ldots\ldots k\}$$

The null hypothesis H_0 will state that the parameters belong to some subspace Θ_0 of the parameter space Θ.

Let $x_1, x_2, \ldots\ldots\ldots\ldots\ldots\ldots x_n$ be a random sample of size $n > 1$

from a population with p.d.f. $f(x, \theta_1, \theta_2 \ldots\ldots\ldots \theta_k)$ where Θ, the parameter space is the totally of all points that $(\theta_1, \theta_2 \ldots\ldots\ldots \theta_k)$ can assume. We want to test the null hypothesis H_0 : $(\theta_1, \theta_2 \ldots\ldots\ldots \theta_k) \in \Theta_0$ against all alternative hypothesis of the type.

$H_1 : (\theta_1, \theta_2 \ldots\ldots\ldots \theta_k) \varepsilon \Theta - \Theta_0$

The likelihood function of the sample observations is given by

$$L = \prod_{i=1}^{n} f(x_i : \theta_1, \theta_2 \ldots\ldots\ldots \theta_k) \qquad \ldots(1)$$

According to the principle of maximum likelihood, the likelihood equation for estimating any parameter θ_i is given by

$$\frac{\partial L}{\partial \theta_i} = 0, (i = 1,2,\ldots k) \quad \ldots\ldots\ldots\ldots(2)$$

Using (2) we can obtain M.L.E for the parameters $(\theta_1, \theta_2 \ldots\ldots\ldots \theta_k)$. Substituting these estimates in (1) we can obtain the maximum values of the likelihood function for variation of parameters in Θ and Θ_0 respectively. Then the criterion for the likelihood ratio is test is defined as the quotient of these two maxima and is given by

$$\lambda = \lambda(x_1, x_2 \ldots\ldots\ldots x_n) = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \frac{\sup_{\theta \in \Theta_0} L(x,\theta)}{\sup_{\theta \in \Theta} L(x,\theta)} \rightarrow \quad (3)$$

Where $L(\hat{\Theta}_0)$ and $L(\hat{\Theta})$ are the maxima of the likelihood

function 1 with respect to the parameters in the regions Θ_0 and Θ respectively.

The quantity λ is a function of the sample observations only and does not involve parameters. Thus λ being a function of the random variables, is also a random variable. Obviously $\lambda > 0$. Further $\Theta_0 \subset \Theta \Rightarrow L(\Theta_0) \leq L(\Theta) \Rightarrow \lambda \leq 1$ Hence we get $0 \leq \lambda \leq 1$

The critical region for testing H_0 (against H_1) is an interval $0 < \lambda < \lambda_0$ ..... (4)

where λ_0 is some number (<1) determined by the distribution of λ and desired probability of type I error. (i.e) λ_0 is given by the equation. $P(\lambda < \lambda_0 | H_0) = \alpha$ .....(5)

For example if g(.) is the p.d.f of λ then λ_0 is determined from the equation

$$\int_0^{\lambda o} g(\lambda \mid H_0) d\lambda = \alpha \qquad \text{..... (6)}$$

A test that has critical region defined in equations (4) and (5) is a likelihood ratio test for testing H_0.

Remarks :

Equations $0 < \lambda < \lambda_0$, and $P(\lambda < \lambda_0 | H_0) = \alpha$ are defined as critical region for testing the hypothesis H_0 by the likelihood ratio test. Suppose the distribution of λ is not known but the distribution of some function of λ is known, then this knowledge can be utilized as given in the following theorem.

Theorem 1.8.(i)

If λ is the likelihood ratio for testing a simple hypothesis H_0 and if $U = \phi(\lambda)$ is a monotonic increasing (decreasing) function of λ then the test based on U is equivalent to the likelihood ratio test. The critical region for the test based on "U is

Proof :

$\phi(0) < U < \phi(\lambda_0)$ $\quad [\phi(\lambda_0) < U < \phi(0)]$

The critical region for the likelihood ratio test is given by

$0 < \lambda < \lambda_0$ where λ_0 is determined by

Let $U = \phi(\lambda)$ be a monotonically increasing function of λ then (i) gives

$$\alpha = \int_{o}^{\lambda_0} g(\lambda \mid H_0) d\lambda \qquad \ldots \qquad (i)$$

$$\alpha = \int_{\phi(0)}^{\phi(\lambda_0)} h(u \mid H_o) du$$

where $h(u|H_0)$ is the p.d.f of U when H_0 is true, and the critical region $0 < \lambda < \lambda_0$ transforms to $\phi(0) < U < \phi(\lambda_0)$.

If $U = \phi(\lambda)$ is a monotonic decreasing function of λ, then the inequalities are reversed and we get the critical region as $\phi(\lambda_0) < U < \phi(o)$.

If we are testing a simple null hypothesis H_0 then there is a unique distribution determined for α. But if H_0 is composite, then the distribution of λ may or may not be unique. In such a case the

distribution of λ may be different for different parameter points in Θ_o and then λ_o is said to be chosen such that

$$\int_{o}^{\lambda o} g(\lambda \mid H_0) d\lambda \leq \alpha.$$

for all values of the parameters in Θo.

If we are dealing with large samples, a fairly satisfactory situation to this testing of hypothesis problem exists as stated in the following theorem.

1.9 PROPERTIES OF LIKELIHOOD RATIO TEST

Likelihood ratio (L.R) test principle is an intuitive one. If we are testing a simple hypothesis H_0 against a simple alternative hypothesis H_1 then the LR principle leads to lemma. This suggests that LR test has some desirable properties, specially large sample properties.

In LR Test, the probability of type I error is controlled by suitably choosing the cut off point λo LR test is generally UMP if an UMP test at all exists. We state below, two asymptotic properties of LR tests.

1. Under certain conditions, $-2 \log_e \lambda$ has an asymptotic chi-square distribution.
2. Under certain assumption, LR test is consistent.

1.10 TEST OF SIGNIFICANCE FOR LARGE SAMPLES

Introduction

There are many problems in which besides estimating the value of a parameter of the population, we must decide whether a statement concerning a parameter is true or false, that is, we must test a hypothesis about a parameter.

To illustrate the general concepts involved in this kind of decisions problems, suppose that a consumer protection agency wants to test a manufacturer's claim that the average life time of electric bulbs produced by him is 200 hours. So it instructs a member of its staff to take 50 electric bulbs from the godown of the company and test them for their lifetime continuously worth the intention of rejecting the claims if the mean life time of the bulbs is below 180 hours(say) otherwise it will accept the claim.

Thus hypothesis is an assumption that we make about an unknown population parameter We can collect sample data from the population, arrive at the sample statistic and then test if the population parameter is true.

Null Hypothesis

A statistical statement about the population parameter assumed before talking the sample for possible rejection on the basis of outcome of sample data is known as a **Null hypothesis**

The null Hypothesis asserts that there is no significant difference between the sample statistic and population parameter.

A hypothesis is said to be **alternative hypothesis** when it is complementary to the null hypothesis.

The null hypothesis and alternative hypothesis is denoted by H_o and H_1 respectively. For example, if we ant to test H_0 that the average height of sailors is 169 cms, then

$H_o : \mu = 169 = \mu_0$ (say)

$H_1 : \mu \neq 169 \neq \mu_0$

Types of Error :

For testing hypothesis, we take a sample from the population, and on the basis of the sample result obtained, we decide whether to accept or reject he hypothesis.

Here, two types of errors are possible. Rejecting H_o, when it is true is called type I error, and the probability of committing type I error is denoted by α.

Accepting H_o when it is false is called type II error and the probability of committing type II error is denoted by β.

Critical region and level of significance :

A region in the sample space which amounts to the rejection of H_o is called the critical region.

Level of significance is the percentage level where the statistical values arrived at random samples falls outside the given range, For example 5% level of significance means 5% of the statistical values arrived at random sample falls outside the given range (a,b) and 95% of values will be within the range (a,b).

Normally in practical problems we consider 5% and 1% level of significance for testing hypothesis.

Test of significance

The tests of significance are

(a) tests of significance for large samples and

(b) tests of significance for small samples.

For large sample size (> 30) all the distributions, like binomial, Poisson etc., are approximated by normal distribution. Thus the normal probability curve can be used for testing hypothesis.

For large samples compute the test statistic 1z1. Then the following results follow.

If $|z| < 1.96$ accept H_0 at 5% level of significance (for two tailed tests)

If $|z| > 1.96$ Reject H_0 at 5% level of significance (for two tailed test)

If $|z| \leq 2.58$ accept H_0 at 1% level of significance (for two tailed test)

If $|z| > 2.58$ reject H_0 at 1% level of significance (for two tailed test)

In a similar fashion readers can learn for single tailed test

Test statistic	Ho	Percentage level of significance	Type of test
$\|z\| < 1.96$	Accept	5%	Two tailed
$\|z\| > 1.96$	Reject	5%	Two tailed

$\|z\| > 2.58$	Reject	1%	Two tailed
$\|z\| \leq 1.64$	Accept	5%	Single tailed
$\|z\| > 1.64$	Reject	5%	Single tailed
$\|z\| \leq 2.33$	Accept	1%	Single tailed
$\|z\| > 2.33$	Reject	1%	Single tailed

Single tailed tests :

Alternative hypothesis H_1 involve single tailed test. Suppose average life of bulbs taken from the population is 1650 hrs.

While testing hypothesis, we assume Ho :$\mu = 1650$. and

$H_1 : \mu \neq 1650$.

Here two possible cases arise, either $H_1 : \mu > 1650$. (or)

$H_1 : \mu < 1650$.

Suppose we are testing any one of the alternative hypothesis, then the test is called **single tailed.**

Procedure for testing of hypothesis

We now summaries below various steps in testing of statistical hypothesis in a systematic manner.

1. Null hypothesis : Set up the null hypothesis H_0.
2. Alternative hypothesis : Set up the alternative hypothesis H_1. This will enable us to decide whether we have to use single tailed (or) two tailed test.
3. Level of significance : Choose the appropriate level of significance.

4. Compute the test statistic.

$$Z=\frac{t-E(t)}{S.E.(t)} \text{ under } H_0.$$

5. conclusion : Compare the calculated value (z) with the tabulated value (z_α).

   If $|z| < z_\alpha$, then accept H_0.

   If $|z| > z_\alpha$, then reject H_0 and accept H_1.

Test of significance for large samples

When the sample size n taken from the population is greater than 30, then it is called large sample.

Under large sample tests, we will see six important test to test the significance.

1. Testing of significance for single mean
2. Testing of significance of difference between two means.
3. Testing of significance for single proportion
4. Testing of significance for difference between two proportions
5. Testing of significance for single standard deviation
6. Testing of significance for difference between two standard derivation.

1.11 TEST OF SIGNIFICANCE FOR SINGLE MEAN :

Let $x_1, x_2, \ldots x_n$ be a random sample of size n from a large population. $X_1, X_2 \ldots X_n$ of size N with mean μ. and variance σ^2. Let $\bar{x}$ be the sample mean.

Step 1 : Set up the null hypothesis $H_0 : \mu = \mu_0$. (i.e) there is no significant difference between population mean and sample mean.

Step 2 : Alternative hypothesis H_1 as $H_1: \mu \neq \mu_0$

$\therefore$ either $H_1 \ \mu < \mu_0$ (or) $H_1 \ \mu > \mu_0$.

Step 3 : Compute z statistic

$$Z = \left| \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \right|$$

If $|z| < z_\alpha$ then accept H_0

If $|z| > z_\alpha$ then reject H_0 and accept H_1 at a level of significance.

Example : 1.11.1

A sample of 400 individuals is found to have a mean height of 67.47 inches. Can it be reasonably regarded as a sample from a large population with mean height of 67.39 inches and standard deviation 1.30 inches?

Solution :

Here $n = 400 \ \bar{x} = 67.47 \ \ \mu = 67.39 \ \sigma = 1.30$

Null hypothesis Let $H_0 : \mu = 67.39$

(i.e) Assume there is significant difference between sample mean and population mean

Alternative Hypothesis $H_1: \mu \neq \mu_0$ (Two tailed test). Now let us compute z statistic

$$|z| = \left| \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \right| = \left| \frac{67.47 - 67.39}{\frac{1.30}{\sqrt{400}}} \right| = 1.23$$

At 5% level of significance, the tabulated value of z_α is 1.96

Since $|z| < z_\alpha = 0.05$, we accept the null hypothesis. (ie) We can arrive conclusion that the sample has been drawn from this population.

Example : 1.11.2

The mean breaking strength of cables supplied by a manufacturer is 1800 with a S.D of 100. By a new technique in the manufacturing process it is claimed that the breaking strength of the cables has increased. In order to test this claim a sample of 50 cables is tested. It is found that the mean breaking strength is 1850. Can we support the claim at 1% level of significance?

Solution :

Here sample size n = 50, sample mean $\bar{x}$=1850, population mean $\mu = 1800$ population standard deviation $\sigma = 100$.

Null hypothesis : $H_0 : \mu = 1800$ (i.e) Assume that mean breaking strength of the cable has not been increased when new technique is applied in the manufacturing process.

Alternative hypothesis : $H_1 : \mu > \mu_1$

(ie.) $H_1 : \mu > 1800$ (Single tailed test) (ie) mean breaking strength of the cable has increased. Compute z statistic as

$$|z| = \left| \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \right| = \left| \frac{1850 - 1800}{\frac{100}{\sqrt{50}}} \right| = \left| \frac{50\sqrt{50}}{100} \right|$$

$|z| = 3.535$

Tabulated value of $z_\alpha = 2.33$ at 1% level of significance (for single tailed test) which is less than calculated value of z. Therefore H_0 is rejected and H_1 is accepted, (ie) we conclude that the mean breaking strength is increased when new technique is applied in the manufacturing process.

Example : 1.11.3

An ambulance service claims that it takes on an average 8.9 minutes to reach its destination in emergency calls. To check on this claim the agency which licenses ambulance services has them timed on 50 emergency calls, getting a mean of 9.3 minutes with a standard deviation of 1.6 minutes. What can they conclude at 5% level of significance??

Solution :

Here $\mu = 8.9$ minutes $n = 50$ $\bar{x} = 9.3$ sample standard deviation $s = 1.6$

Null hypothesis $H_0 : \mu = 8.9$ minutes

Alternative hypothesis $H_0 : \mu \neq 8.9$

Test statistic

$$|z| = \left| \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \right| = \left| \frac{9.3 - 8.9}{\frac{12.6}{\sqrt{50}}} \right|$$

$|z| = 1.768$

Tabulated value for z at 5% level of significance for large samples is 1.96

Since $|z| < z_{0.05}$, the null hypothesis is accepted and we conclude that the average time taken to reach the destination is 8.9 minutes.

Example: 1.11.4

A sample of 900 members has a mean 3.5 cms and S.D. 2.61 cms. Is the sample from a large population of mean 3.25 sms and S.D 2.61 cms?

If the population is normal and its mean is unknown, find 95% and 98% fiducial limits of true mean.

Solution :

Here n = 900 $\bar{x}$ = 3.5 s = 2.61 $\mu = 3.25, \sigma = 2.61$

Null hypothesis H_0: $\mu = 3.25$

Alternative hypothesis H_1 : $\mu \neq 3.25$ (Two – tailed)

Test statistic

$$|z| = \left| \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \right| = \left| \frac{3.5 - 3.25}{\frac{2.61}{\sqrt{900}}} \right|$$

$$= \left| \frac{0.25}{\frac{2.61}{30}} \right| = 2.8735$$

Since $|z| > 1.96$, we conclude that H_0 is rejected at 5% level of significance.

95% fiducial limits for μ are $\bar{x} \pm 1.96\, \sigma / \sqrt{n}$

$$\Rightarrow 3.5 \pm 1.96 \times \frac{2.61}{\sqrt{900}}$$

$$\Rightarrow 3.5 \pm 0.1705$$

(i.e) 3.32.95 and 3.6 705

98% fiducially limits for μ are given by

$$\bar{x} \pm 2.33 \frac{\sigma}{\sqrt{n}}$$

$$\Rightarrow 3.5 \pm 2.33 \times \frac{2.61}{\sqrt{900}}$$

$$\Rightarrow 3.5 \pm 0.2027$$

$$\Rightarrow 3.2973 \text{ and } 3.7027.$$

Remark 1

The minimum sample size for estimating μ with confidence co-efficient $(1-\alpha)$ is given by $\sigma^2 Z_\alpha^2 / E^2$, where z_α is the significant value of z at level of significance α and E is the permissible error in the estimate.

Remark 2:

The minimum sample size for estimating population proportion P with confidence co-efficient $(1-\alpha)$ is given by

$n = PQZ_\alpha^2 / E^2$.

Example : 1.11.5

The mean muscular endurance score of a random sample of 60 subjects was found to be 145 with a S.D. of 40. Construct a 95% confidence interval for the true mean. Assume the sample size to be large enough for normal approximation. What size of sample is required to estimate the mean within 5 of the true mean with a 95% confidence?

Solution :

We are given n = 60 $\bar{x}$ = 145, s = 40

95% confidence is interval for population mean are $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} = 145 \pm 1.96 \frac{40}{\sqrt{60}}$

(i.e) 134.88 and 155.12

Also $N = \left(\frac{1.96\sigma}{E}\right)^2 \Rightarrow \left(\frac{1.96x40}{5}\right)^2 = 246.$

1. 12 TEST OF SIGNIFICANCE FOR DIFFERENCE OF MEANS

Let $\bar{x}_1$ be the mean of sample of size n_1 from a population with mean μ_1 and variance σ_1^2 Let $\bar{x}_2$ be the mean of sample of size n_2 from a population with mean μ_2 and variance σ_2^2. Since sample sizes are large, $\bar{x}_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$ and $\bar{x}_2 \sim N(\mu_1, \frac{\sigma_2^2}{n_2})$

$$\text{Also } z = \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{S.E(\bar{x}_1 - \bar{x}_2)} \sim N(0,1)$$

Under null hypothesis $H_0 : \mu_1 = \mu_2$

$$E(\bar{x}_1 - \bar{x}_2) = E((\bar{x}_1) - E((\bar{x}_2) = \mu_1 - \mu_2 = 0$$

$$V(\bar{x}_1 - \bar{x}_2) = V((\bar{x}_1) + V((\bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

(Since the covariance terms vanish since $\bar{x}_1$, $\bar{x}_2$ are independent)

Hence under $H_0 : \mu_1 = \mu_2$ The test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad \rightarrow (1)$$

Note 1 :

Under $H_0 : \mu_1 = \mu_2$ if the samples are drawn from the same population where

$\sigma_1 = \sigma_2 = \sigma$, then

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \rightarrow (2)$$

Note 2:

If σ_1 and σ_2 are not known and $\sigma_1 \neq \sigma_2$, the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad \rightarrow (3)$$

Note 3 :

In equation (2) if σ is not known, and $\sigma_1 = \sigma_2$. We use $\sigma^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$ to calculate σ and use equation (2), we have

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 n_2}}}$$

Example : 1.12.1

The electric light bulbs of manufacturer A, have mean life time of 1600 hours with a S.D of 200 hours, while those of manufacturer B have a mean life time of 1400h with a S.D of 100 hr. If random samples of 125 bulbs from each manufacturer are selected then what is the probability that the bulbs will have a mean lifetime that is at least (a) 160 and (b) 250h more than B bulbs?

Solution :

We are given

$$\mu_{\bar{x}_A - \bar{x}_B} = \mu_{\bar{x}_A} - \mu_{\bar{x}_B} = 1600 - 1400 = 200$$

$$\sigma_{\bar{x}_A - \bar{x}_B} = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} = \sqrt{\frac{200^2}{125} + \frac{100^2}{125}} = 20 \text{ h.}$$

Give $n_A = n_B = 125$,

$\mu_{\bar{x}_A} = 1600$, $\mu_{\bar{x}_B} = 1400$, $\sigma_A = 200h, \sigma_B = 100h$

$$Z = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_{\bar{x}_A} - \mu_{\bar{x}_B})}{\sigma_{\bar{x}_A - \bar{x}_B}}$$

$$= \frac{(\bar{x}_A - \bar{x}_B) - 200}{20}$$

(a) If $\bar{x}_A - \bar{x}_B = 160$ h, then

$$Z = \frac{160 - 200}{20} = -2$$

$$P(\bar{x}_A - \bar{x}_B \geq 160) = P(z \geq -2)$$

$$= 0.5 + 0.4772$$

$$= 0.9772$$

b. If $\bar{x}_A - \bar{x}_B = 250$ h, then

$$P(\bar{x}_A - \bar{x}_B \geq 250) = P(z \geq 2.5)$$

$$= 0.5 - 0.4938$$

$$= 0.0062.$$

Example : 1.12.2

Samples of students were drawn from two universities in kgm, means and standard deviations are calculated. Make large sample test to the significance of the difference between the means.

	Mean	S.D.	Sample size
University A	55	10	400
University B	57	15	100

Solution :

Here are given,

$n_1 = 400$ $\quad \bar{x} = 55$ $\quad s_1 = 10$

$n_2 = 100$ $\quad \bar{x}_2 = 57 s_2 = 15$

Null hypothesis : $H_o : \mu_1 = \mu_2$

(i.e) There is no significant difference between two sample means.

Alternative hypothesis : $H_1 : \mu_1 \neq \mu_2$

Level of significance : $\alpha = 5\%$

Compute test statistic

$$|z| = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0.1)$$

$$= \frac{55 - 57}{\sqrt{\frac{10^2}{400} + \frac{15^2}{100}}} = |-1.2649|$$

Since $|z| < 1.96$, we accept H_0 at 5% level of significance (i.e) the difference between the two sample means not significant.

Example : 1.12.3

The average hourly wage of a sample of 150 workers in a plant A was Rs. 2.56 with a S.D. of Rs. 1.08. The average wages of a sample of 200 workers in a plant B was 2.87 with a S.D of Rs. 1.28. Can an applicant safely assume that the hourly wages paid by plant B are lower than those paid by plant A? use 1% level of significance.

Solution :

Let X_1 and X_2 the hourly wages (in Rs) of works in plant A and plant B respectively. We are given $n_1 = 150$, $\bar{x}_1 = 2.56$, $s_1 = 1.08$ $n_2 = 200$, $\bar{x}_2 = 2.87$ $s_2 = 1.28$

Null hypothesis :

H_o: $\mu_1 = \mu_2$ (ie) there is no significance difference between the mean level of wages of workers in plant A and B.

Alternative hypothesis : H_1 : $\mu_1 \neq \mu_2$ (i.e) H_1: $\mu_1 > \mu_2$

Test statistic :

$$Z = \left| \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \right|$$

$$= \left| \frac{2.56 - 2.87}{\sqrt{\frac{(1.08)^2}{150} + \frac{(1.28)^2}{200}}} \right|$$

$$= +2.46$$

Tabulated value of z at 1% level of significance for single tailed test is 2.33, and since $|z| > 2.33$, the null hypothesis is rejected and we conclude that the hourly wages paid by the plant B are lower than those paid by plant A.

Example : 1.12.4

Two samples of cables of sizes each 100 are taken in order to test the breaking strength and their mean and standard deviations are given.

Cable I	Cable II
$\bar{x}_1 = 1925$	$\bar{x}_2 = 1905$
$\sigma_1 = 40$	$\sigma_2 = 30$

Test whether mean breaking strength of the two cables are equal at 1% level of significance.

Solution : Null hypothesis $H_o : \mu_1 = \mu_2$

(i.e) there is no significant difference between the mean breaking strength of two cables.

Alternative hypothesis : $H_1: \mu_1 \neq \mu_2$ (two tailed)

Level of significance: $\alpha = 0.01$

Computation of test statistic :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \qquad Z = \frac{1925 - 1905}{\sqrt{\frac{40^2}{100} + \frac{30^2}{100}}} = 4$$

Tabulated value of z at 1% level of significance for two tailed test is 2.58. Since $|z| > 2.58$, the null hypothesis is rejected and we conclude that the mean breaking strength of the cables differ significantly.

Example : 1.12.5

A random sample of 200 villages from Coimbatore district gives mean population per village at 485 with a S.D of 50. Another random sample of the same size from the same district gives the mean population per village at 510 with a S.D of 40. Is the difference between the mean values given by the two samples statistically significant? Justify your answer. Assume σ is not known.

Solution :

We are given

$n_1 = 200 \quad \bar{x}_1 = 485, \; s_1 = 50$

$n_2 = 200 \quad \bar{x}_2 = 510, \; s_1 = 40$

Solution

Both samples are taken from the same population and therefore $\sigma_1 = \sigma_2 = \sigma$ and σ is known.

Nul Hypothesis :

H_0: $\mu_2 = \mu_2$ (i.e) difference between the means of the samples is not significant.

Alternative hypothesis ; $H_1 : \mu_2 \neq \mu_2$ (two tailed)

Level of significance : $\alpha = 0.05$

Test Statistic

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_2} + \frac{s_2^2}{n_1}}}$$

$$= \frac{485 - 510}{\sqrt{\frac{50^2}{200} + \frac{40^2}{200}}}$$

$$= -5.5216$$

$|z| = 5.52$ and tabulated value of z at 1% level of significance is 2.58

1.13 TEST OF SIGNIFICANCE FOR SINGLE PROPORTION

Let X be number of success in n independent trails with constant probability P of success for each trail, we know E(x) = np and var(X) = nPQ, where Q = 1-P

Sample proportion of success = $\frac{X}{n} = p(say)$

$$E(p) = E\left(\frac{X}{n}\right) = \left(\frac{1}{n}\right) E(X) = \left(\frac{1}{n}\right) nP = P$$

ie $E(p) = P$ and var $(p) = V\left(\frac{X}{n}\right) = \frac{1}{n^2} \text{Var}(X) = \frac{1}{n^2} nPQ = \frac{PQ}{n}$

$$\therefore \text{S.E}(p) = \sqrt{\frac{PQ}{n}}$$

Therefore $Z = \frac{p - E(p)}{S.E.(p)} = \frac{p - (P)}{\sqrt{\frac{PQ}{n}}} \sim N(0.1)$

Step 1 : Under Null hypothesis set H_o: $P = P_o$

(ie) there is no significance difference between the

population proportion of success and sample proportion of success.

Step 2 : Alternative Hypothesis : $H_1 : P \neq P_0$

Step 3 : Level of significance : Choose appropriate level of significance

Step 4 : Compute test statistic

$$Z = \frac{p-(P)}{\sqrt{\frac{PQ}{n}}}$$

Step 5 : If $|z| < z_\alpha$, accept H_0 at α% level of significance, otherwise reject H_0 and accept H_1.

To find confidence limits

Note 1:

Suppose the sample is taken from a finite population of size N, then

$$\text{S.E (p)} = \sqrt{\left(\frac{N-n}{N-1}\right) \bullet \frac{PQ}{n}}$$

Note 2 : The probable limits for a normal variate X are $E(X) \pm 3\sqrt{V(X)}$

Note 3: The probable limits for the sample proportion of successes are $P \pm 3\sqrt{PQ/n}$

Note 4 : If P is not known, taking p (sample proportion) as an estimate of P, the probable limits for the proportion in the population are $p \pm 3\sqrt{pq/n}$.

Note 5: 95% confidence limits for P are given $p \pm 1.96 \sqrt{pq/n}$

Note 6: 99% confidence limits for P are given by $p \pm 2.58 \sqrt{pq/n}$

Example : 1.13.1

A random sample of 400 persons were taken from the large population and 120 were found to be females. Can we say men and women are in the ratio 5:3 in the population? (Use 1% level of significance)

Solution :

We are given n = 400, X = no of females in the sample = 120

$$p = \text{sample proportion of females} = \frac{120}{400} = 0.30$$

Null hypothesis : $H_0 : P = \frac{3}{8} = 0.375$

(ie) population proportion of females is 0.375

Alternative hypothesis : $H_1 : P \neq 0.375$ (Two tailed)

Level of Significance : $\alpha = 1\%$

Computation of Z statistic is

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$$

$$Z = \frac{0.300 - 0.375}{\sqrt{\frac{0.375x0.625}{400}}}$$

$$= -3,125$$

$$|z| = 3.125$$

Tabulated vale of z at 1% level of significance is 2.58. Since $|z| > Z_\alpha$, the null hypothesis is rejected and we conclude that the males and females in the population are not in he ratio 5:3.

Example : 1.13.2

A manufacturer claims that only 6% of his products supplied by him are defective. A random sample of 800 products contained 40 defectives. Test the claim of the manufacturer to be rejected?

Solution :

P = proportion of defective products in the sample

$$= \frac{40}{800} = 0.05$$

P = proportion of defective products in the population

$= 0.06$

Null Hypothesis : Ho : P = 0.06 is true

Alternative Hypothesis : H_1 : P > 0.06

Level of significance : $\alpha = 5\%$

Computation of z statistic

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} = \frac{0.05 - 0.06}{\sqrt{\frac{0.06 \times 0.04}{800}}} = \frac{-0.01}{0.0017} = -5.80.$$

Since |z| = 5.8 > 1.96, we reject the null hypothesis at 5% level of significance and we conclude that manufacture's claim is not acceptable.

Example : 1.13.3

In a sample of 2000 people in TamilNadu 1080 are vegetarians and the rest are non vegetarians. Can we assume that both vegetarians and non vegetarians are equally popular in this state at 1% level of significance?

Solution : We are given n = 2000

X = No of vegetarians = 1080

p = Sample proportion of vegetarians = $\frac{1080}{2000} = 0.54$

Null Hypothesis H_0: P = 0.5

(ie.) Both vegetarians and non vegetarians are equally popular in the state.

Alternative Hypothesis H_1: P $\neq$ 0.5 (two tailed)

Test statistic : Under Ho test statistic is

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} \sim N(0,1) \text{ (Since n is large).}$$

$$Z = \frac{0.54 - 0.5}{\sqrt{\frac{0.5x0.5}{2000}}} = \frac{+0.04}{0.0111} = 3.60$$

Conclusion : The significant (or) critical value of Z at 1% level of significance for two tailed test is 2.58. Since |z| > 2.58, the hull hypothesis is rejected and we conclude that vegetarians and non vegetarians are not equally popular in Taminadu State.

Example : 1.13.4

A sample of 900 days was taken for metrological records of a certain district and 100 of them were found to be foggy. What are the possible limits to the percentage of foggy days in the district.?

Solution :

Hence n = 900

p = Proportion of fogy days in the sample

$$= \frac{100}{900} = 0.11$$

q = 1 - 0.11 = 0.89

The limits for the population proportion P are given by

$$p \pm 3\sqrt{\frac{pq}{n}} = 0.11 \pm 3\sqrt{\frac{0.11x0.89}{900}}$$

$$= 0.11 \pm 3 \times 0.01014296$$

$$= 0.11 \pm 0.031288$$

(i.e)0.078712 and 0.141288 (ie.) 8% and 14.12%

1.14 TEST OF SIGNIFICANT DIFFERENCE BETWEEN TWO SAMPLE PROPORTIONS

Let us consider two samples of sizes n_1 and n_2 drawn from two different populations. Let p_1 and p_2 are the two sample proportions.

Under null hypothesis, we have

H_0: $P_1 = P_2$ (i.e) there is no significant difference between two sample proportions.

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{where } P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \text{ and } Q = 1 - P$$

Example : 1.14.1

In a sample of 600 students 400 are found to use pens. In another college, from a sample of 900 students, 450 were found to use pens. Test whether the two colleges are significantly different with regard to using pens.

Solution :

$n_1 = 600$ $\qquad$ $n_2 = 900$

$x_1 = 400$ $\qquad$ $x_2 = 450$

p_1 = proportion of pens used from the first sample.

$$= \frac{x_1}{n_1} = \frac{400}{600} = 0.66$$

p_2 = proportion of pens used from the second sample.

$$= \frac{x_2}{n_2} = \frac{450}{900} = 0.5$$

Ho: $P_1 = P_2$ (i.e) thee is no difference between the proportion of pens used in two colleges.

H_1: $P_1 \neq P_2$ (two tailed test)

Level of significance : $\alpha = 5\%$

Computation of Z statistic

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where $P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

$$= \frac{x_1 + x_2}{n_1 + n_2} = \frac{400 + 450}{600 + 900} = \frac{850}{1500} = 0.5666$$

and $Q = 1 - 0.5666 = 0.4334$

$$\therefore Z = \frac{0.66 - 0.5}{\sqrt{0.5666 \times 0.4334 \left(\frac{1}{600} + \frac{1}{900} \right)}}$$

$$= \frac{0.16}{0.02574} = 6.22$$

Tabulated value of z at 5% level of significance is 1.96 (for two tailed test)

Since $|z| > 1.96$, the null hypothesis is rejected and we conclude that, with regard to the usage of pens, there is significant difference exists between colleges.

Example : 1.14.2

500 articles from a factory are examined and found to be 2% defective. 800 similar articles from a second factory are found to have only 1.5% defective. Can it reasonably be concluded that the product of the first factory are inferior to those of second?

Solution

We are given, $n_1 = 500$, $n_2 = 800$

P_1 = proportion of bad articles in the first population = 0.02

P_2 = proportion of bad articles in the second population = 0.15

$Q_1 = 1 - 0.02 = 0.98$

$Q_2 = 1 - 0.015 = 0.985$

Ho : Sample proportions are equal

(i.e) $H_0 : P_1 = P_s$

Alternative hypothesis $H_1: P_1 < P_2$ (Left tailed)

Test statistic

$$Z = \frac{P_1 - P_2}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}}$$

$$= \frac{0.02 - 0.015}{\sqrt{\frac{0.02x0.98}{500} + \frac{0.015x0.985}{800}}} = \frac{0.005}{0.0075} = 0.666 = 0.67$$

$\Rightarrow |z| = 0.67.$

Conclusion

Tabulated value of z at 5% level of significance for single tailed test is 1.645.

Since $|z| < 1.645$, H_0 is accepted and we conclude that products do not differ in quality in the two samples.

Example : 1.14.3

A manufacturing firm claims that its brand A product outsells its brand B product by 10%. If it is found that 50 out of a sample of 200, persons prefer brand A and 20 out of another sample of 100 persons prefer brand B. test whether the 8 difference is a valid claim.

Solution :

Here $n_1 = 200$, $n_2 = 100$,

$$P_1 = \frac{50}{200} = 0.25 \text{ and } P_2 = \frac{20}{100} = 0.2$$

Also $P_1 - P_2 = 10\% = 0.10$

Null Hypothesis H_0 : Assume that the difference is a valid claim (i.e) $P_1 - P_2 = 0.10$

Alternative Hypothesis H_1 : $P_1 - P_2 \neq 0.10$ (two tailed test)

Under Ho test statistic is

$$Z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

Here $\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{50 + 20}{200 + 100} = 0.233$ and

$\hat{Q} = 1 - \hat{P} = 1 - 0.233 = 0.767$

$$\therefore Z = \frac{(0.25 - 0.2) - 0.10}{\sqrt{0.233 \times 0.767\left(\frac{1}{200} + \frac{1}{100}\right)}}$$

$$= \frac{0.05 - 0.10}{0.0517} = -0.967$$

$$\Rightarrow |z| = 0.967$$

Since $|z| < 1.96$ at 5% level of significance null hypothesis H_0 is accepted and we conclude that the difference is a valid claim.

Example : 1.14.4

In a random sample of 500 students the university teaching departments, it was forward that 400 students passed in examination. In another sample of 600 students of the affiliated colleges, the number of passes in the same examination was found to be 400. Find out whether proportion of passes in the university teaching departments is significantly lesser than the passes in the university teaching departments and affiliated colleges taken together.

Solution :

Here we are given $n_1 = 500$, $n_2 = 600$

$p_1 = \frac{400}{500} = \frac{4}{5} = 0.8$, $p_2 = \frac{400}{600} = 0.66$ and $q_1 = 1 - p_1 = 0.2$

Null hypothesis Ho : p_1 and $\hat{p}$ where $\hat{p}$ is the pooled estimate (i.e) proportion of passes

in the university teaching departments and affiliated colleges taken together do not

differ significantly.

$$SE(\hat{p} - p_1) = \sqrt{\frac{\hat{p}\hat{q}}{n_1 + n_2} \times \frac{n_2}{n_1}}$$

$$\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{500 \times 0.8 + 600 \times 0.66}{500 + 600}$$

$$= \frac{400 + 400}{1100} = 0.0727 = 0.73$$

and $\hat{q} = 1\text{-}0.73 = 0.27$

$$\therefore \text{S.E}(\hat{p}-p_1) = \sqrt{\frac{0.73\times 0.27}{500+600}\times\frac{600}{500}} = \sqrt{\frac{118.26}{550000}} = 0.0146$$

Test Statistic :Under null hypothesis H_0, the test statistic is

$$Z = \frac{\hat{p}-p_1}{S.E(\hat{p}-p_1)} \sim N(0,1) = \frac{0.73-0.8}{0.0146} = -4.79$$

$\Rightarrow |z| = 4.79$

Conclusion : Since calculated value of |z| is greater than 3 it is highly significant.

Hence null hypothesis : H_0 is rejected and we conclude that there is significant difference between p_1 and $\hat{p}$.

Exercise

1 In a year there are 956 deaths in a town A of which 52.5% were males while in town A and B combined, this proportion in a total of 1406 deaths was 0.496. Is there any significant difference in the proportion of male deaths in the two towns?

(Ans : H_0 is rejected, $|z| = 3.368$)

1. Explain the procedure for testing the significant difference between the two sample proportion for large sample.
2. In two large populations, there are 30 and 25 percent respectively of blue – eyed people. Is this difference likely to be hidden in samples of 1200 and 900 respectively from the two populations? (Use 1% level of significance)

(Ans : $|z| = 2.56$, H_0 is accepted)

3. On the basis of their total scores 400 candidates of a civil service examination are divided into two groups the upper 40% and the remaining 60%. Consider the first question of this examination. Among the first group 100 had the correct answer and 200 had correct answer in the second group. On the basis of these result, can one conclude that the first question is good at discriminating ability of the type being examined here?

(Hint : $n_1 = \frac{40}{100} \times 400 = 160$, $n_2 = \frac{60}{100} \times 400 = 240$

$|z| = 4.96$, H_0 is rejected, Highly significant)

4. A cigarette manufacturing firm claims that its brand A of the cigarettes outsells its brand B by 8%. If it is found that

the cigarettes outsells its brand B by 8%. If it is found that 42 out of a sample for 200 smokers prepare brand A and 18 out of another random sample of 100 smokers prefer brand B , test whether the 8% difference is a valid claim. (Use 5% level significance) (Ans : $|z| = 1.02$, H_0 is accepted that is a difference of 8% in the sale of two brands of cigarettes is a valid claim)

1.15 TEST OF SIGNIFICANCE OF THE DIFFERENCE BETWEEN TWO SAMPLE STANDARD DEVIATIONS

Procdure:

If s_1, s_2 are the standard deviations of two independent samples then under null hypothesis H_0: $\sigma_1 = \sigma_2$, (i.e.) there is no significance difference between two sample standard deviations.

The test statistic is

$$z = \frac{s_1 - s_2}{S.E(s_1 - s_2)} \sim N(0,1) \text{ for large samples}$$

$$= \frac{s_1 - s_2}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_1^2}{2n_2}}} \quad \text{if } \sigma_1, \sigma_2 \text{ are known or}$$

$$z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_1^2}{2n_2}}} \quad \text{if } \sigma_1, \sigma_2 \text{ are not known}$$

If the calculated value of $|z|$ is less than the tabulated value

1.96 (or 2.58) then we accept H_0 at 5% (or 1%) level of significance for two tailed test respectively, otherwise we reject H_0 at these level significances.

Example : 1.15.1

For two samples the following data are given

$n_1 = 1000 \qquad \bar{x}_1 = 67.42 \qquad s_1 = 2.58$

$n_2 = 1200 \qquad \bar{x}_2 = 67.25 \qquad s_2 = 2.50$

Is the difference between the standard deviations significant?

Solution:

Null hypothesis $H_0 : \sigma_1 = \sigma_2$ (i.e) there is no significant difference between sample standard deviations.

We are given $n_1 = 1000 \; \bar{x}_1 = 67.42 \qquad s_1 = 2.58$

$n_2 = 1200 \qquad \bar{x}_2 = 67.25 \qquad s_2 = 2.50$

Test statistic

$$z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_1^2}{2n_2}}} = \frac{2.58 - 2.50}{\sqrt{\frac{(2.58)^2}{2(1000)} + \frac{(2.50)^2}{2(2100)}}}$$

$$= \frac{0.08}{0.07702} = 1.029$$

Since $|z| = 1.03 < 1.96$ we accept H_0 at 5% level of significance and we conclude that there is no significant difference between two sample standard deviations.

Example : 1.15.2

Two sets of 100 students each were taught to read by two different methods. After the instruction were over, a reading test given to them reveal $\overline{x_1} = 73.4$ $\overline{x_2} = 70.3$

$s_1 = 8$, $s_2 = 10$, $H_0 : \sigma_1 = \sigma_2$

Solution: $n_1 = 100$ $\overline{x_1} = 73.4$ $s_1 = 8$

$n_2 = 100$ $\overline{x_2} = 70.3$ $s_2 = 10$

Null hypothesis $H_0 : \sigma_1 = \sigma_2$

Test Statistic

$$z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_1^2}{2n_2}}} = \frac{8-10}{\sqrt{\frac{64}{2(100)} + \frac{100}{2(100)}}} = \frac{-2}{0.9055} = -$$

2.21

Conclusion :

Since $|z| = 2.21 > 1.96$ we reject H_0 at 5% level of significance and we conclude that, there is a significant difference between two sample standard deviations.

Exercise

1. Explain the procedure of test of significance of two sample standard deviations for large samples.

2. In questions 2 and 3 test the significance of the difference between the two sample SD's for the following data at 5% level

Sample	Size	Mean	S.D
I	100	61	4
II	200	63	6

(Ans: $|z|$ - 4.8508, reject H_0)

3. | | | | |
|---|---|---|---|
| Sample A | 250 | 120 | 12 |
| Sample B | 400 | 124 | 14 |

(Ans: $|z|$ - 2.73, reject H_0)

4. A random sample of 2400 men from coimbatore city gives the mean pay Rs. 800 per week with a S.D of Rs. 120 and a random sample of 2000 men from Chennai gives the mean pay of Rs. 1000 per week with a S.D of Rs. 160. Discuss, whether standard deviation levels of pay of men from the two cities differ significantly.

(Ans: $|z|$ -13.05, reject H_0)

5. Intelligent tests were given to two groups of boys and girls.

	Size	Mean	S.D
Girls	60	75	8
Boys	100	73	10

Examine if the difference between S.D scores is significant. (Ans: $|z|$ - 1.967, reject H_0)

CHAPTER II

TEST OF HYPOTHESIS FOR SMALL SAMPLES

t- DISTRIBUTION

To establish confidence interval at certain level of significance for the population mean when the standard derivation is unknown and to find confidence interval for population mean when the sample size is small, we use 't' distribution discovered by William Gosset under the nickname 'Student'. Hence, it is called "student's t-distribution.

2.1 TEST OF SIGNIFICANCE FOR SMALL SAMPLES–

t TEST SMALL SAMPLE TESTS

In this chapter we study the test of significance for small samples of size < 30. In all exact sample tests, we assume that the population from which the samples drawn are normal.

The 't' test of significance (for small sample) student's t:

Definition :

Let $x_1, x_2 \ldots\ldots\ldots x_n$ be a random sample of size n from a normal population with mean μ and variance σ^2. The student's t test is defined in the statistic as $t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}}$

where $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$ is the sample mean and $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is an unbiased estimate of the population variance σ^2 and 't' follows students t distribution with $v = n - 1$ degrees of freedom with p.d.f $f(t) = (1+t^2/v)^{\frac{-v+1}{2}}$, where v = n-1 and y_o is a constant

determined by $\int_{-\infty}^{\infty} f(t)dt = 1$.

This is known as student's 't' distribution.

Note : When sample standard deviation s is given, then student's test statistic

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}$$

Properties of student's 't' distribution

1. The variable ranges from - ∞ to ∞ like normal distribution.
2. It is asymmetrical distribution and its mean is zero.
3. 't' distribution has more deviation than standard normal distribution
4. The variance of the distribution is greater than 1. but tends to 1 as $n \to \infty$. When $n \to \infty$ the distribution becomes normal
5. variance = $\frac{v}{v-2}$ if $v > 2$ where $v = n-1$.

Characteristics of the t – distribution

1. t is also symmetrical like normal distribution . Its values vary from o to ∞
2. The curve of the t distribution is flatter than the normal distribution
3. Number of degrees freedom for this distribution is n-1
4. It is unimodal distribution

5. It has no mean for $n = 2$ (or $v = n - 1 = 1$) and no variance for $n \leq 3$ and $v \leq 2$. However for $v > 1$ its mean

$$\mu = E(t) = 0 \text{ and for } v > 2, \text{ its variance } v(t) = \frac{v}{v-2}$$

6. The variance of 't' distribution approaches 1 as the sample size $n \to \infty$. In general for $n \geq 30$, its variance is approximately the same as that of z distribution.

In the following figure we are given two t distribution curves, one for $v = 5$ and other $v = 20$, drawn against standard normal distribution curve.

Applications of 't' distributions

The 't' distribution has many applications, some of which are given below :

1. To test the significant difference between sample mean $\bar{x}$ and hypothetical (population) mean μ.
2. To test the significance difference between two sample means.
3. To test the significance difference between observed sample correlation co-efficient and sample regression co-efficient.
4. To test the significance of observed partial and multiple correlation co-efficient.

't' test for single mean :

Suppose we want to test :

i. if a random sample $x_1, x_2 \ldots x_n$ of size n has been drawn from a normal population with a specified mean, say μ_0

ii. If the sample mean differs significantly from the hypothetical value μ_0 of the population mean.

Under the null hypothesis H_0

i. The sample has been drawn from the population with mean μ (or)

ii. There is no significant difference between the sample mean $\bar{x}$ and the population mean μ.

The test statistic

$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} \text{ where } \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ and } S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

follows 't' distribution with (n-1) degrees of freedom.

Conclusion :

If $|t|<$ tabulated value at (n-1) d.f, the null hypothesis H_0 is accepted. Otherwise H_0 is rejected.

Example : 2.1.1

A type of insecticide is filled in bags using a machine. A random sample of 10 bags were selected and their weights (in kgs) are given below.

50 49 52 44 45 48 46 45 49 45 Can we assume this sample be taken from a population of mean 50?

Solution :

Null Hypothesis : $H_0 : \mu = 50$ (kgs) average weight of the filled bags = 50 kgs

Alternative hypothesis $H_1: \mu \neq 50$ kgs (two tailed test)

Level of significance Let $\alpha = 5\%$

X	$(x-\bar{x})^2$
50	7.29
49	2.89
52	22.09
44	10.89
45	5.29
48	0.49
46	1.69
45	5.29
49	2.89
45	5.29
473	64.1

$$\bar{x} = \frac{473}{10} = 47.3 \text{ and } S^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2 = \frac{64.1}{10-1} = 7.12$$

't' test statistic $$t = \frac{\bar{x} - \mu}{\sqrt{S^2/n}} = \frac{47.3 - 50}{\sqrt{\frac{7.12}{10}}} = -3.2$$

$$|t| = 3.2$$

Tabulated value of t at (n-1=10-1 = 9) d.f from the t table for 5% level of significance is 2.262 Since calculated value of t > tabulated value, the null hypothesis H_0 is rejected and we conclude that average weight of the bag is not 50 kgs.

Example : 2.1.2

A soap manufacturing soap sells soaps through retail shops. Before advertisement, the average sale of soap in a shop is 140 dozens. After advertisement, a random sample of 26 shops is taken and the mean and S.D. is 147 dozens and 16 respectively. Can we say, whether the advertisement is successful?

Solution :

We are given n = 26, $\bar{x}$ = 146, dozens s = 16

Null Hypothesis : Ho : μ = 140 dozens

(ie) there is no increase in sale of soaps by advertisement

Alternative hypothesis : H_1:μ > 40 doznes (right tailed)

Level of significance : Let α = 5%

Computation of 't' statistic

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} = \frac{147-140}{16/\sqrt{25}} = 2.19$$

Tabulated value of t for 26 - 1= 25 degrees of freedom at 1% level of significant for single tailed test is 2.060. Since |t| > 2.06, H_0 is rejected and we conclude that sales of soap would be increased trough advertisement.

Example : 2.1.3

Ten individuals are chosen at random from a normal population and their heights are found to be 63,63,66,67,68,69,70,70,71,71 inches. Test if the sample belongs to the population whose mean heights is 66" (given $t_{0.05} = 2.62$ for 9 d.f)

Solution :

We are given n = 10, $\bar{x} \frac{\sum x}{n} = \frac{678}{10} = 67.8$

X	$(x-\bar{x})^2$
63	23.04
63	23.04
66	3.24
67	0.64
68	0.04
69	1.44
70	4.84
70	4.84
71	10.24
71	10.24
678	81.6

$$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2 = \frac{1}{9}(81.6) = 9.066$$

Null hypothesis : $H_0 : \mu = 66$"

Alternative hypothesis $H_1: \mu \neq 66$"

Level of significance : $\alpha = 5\%$

Test of statistic

$$t = \frac{\bar{x} - \mu}{\sqrt{S^2/n}} = \frac{67.8 - 66}{\sqrt{9.066/10}} = \frac{1.8}{0.95215} = 1.89$$

tabulated value of t at 5% level of significance for 9 d.f is 2.62 Since $|t| < 2.62$, H_0 is accepted and is concluded that the sample has been drawn from the population whose mean height is 66".

Exercise

1. The mean weekly sales of soap bars in departmental stores was 146.3 bars per store. After an advertising campaign the mean weekly sales in 22 stores for a typical week increased to 153.7 and showed a standard deviation of 17.2. Was the advertising campaign successful? (Ans : $|t| = 9.03$, H_0 is rejected)
2. Explain the applications of 't' distributions.
3. Explain the properties of student's 't' distribution.
4. A random sample of 10 days had the following I.Q's 70,120,110,101,88,83,95,98,107,100. Do these data support the assumption of a population mean I.Q of 100?

2.2: PAIRED 't' TEST FOR SINGLE MEAN

It the pairs of values are in some way associated, we have to find the differences of the associated pairs and apply for single

mean (ie) $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n-1}}$ to test if the means of the differences is significantly different from zero. Then the degree of freedom is n - 1.

Here the test statistic is

$$t = \frac{\bar{d}}{s/\sqrt{n-1}} \text{ (or) } \frac{\bar{d}}{S/\sqrt{n}}$$

where $\bar{d}$ is the mean of the paired differences (ir.) $d_i = x_1 - y_i$ and $\bar{d} = \bar{x} - \bar{y}$ where (x_i, y_i) are the paired data i = 1,2, ...n.

Example : 2.2.1

The scores of 10 candidates prior and after training are given below

Prior :	84	48	36	37	54	69	83	96	90	65
After:	90	58	56	49	62	81	84	86	84	75

Solution :

Null Hypothesis H_0 : $\mu_1 = \mu_2$: the training is not effective

Null Hypothesis H_1 : $\mu_1 \neq \mu_2$: the training is effective

Total										
x_1:	84	48	36	37	54	69	83	96	90	65
x_2:	90	58	56	49	62	81	84	86	84	75
$d = x_1 - x_2$	-6	-10	-20	-12	-8	-12	-1	10	6	-10	-63
d^2	36	100	400	144	64	144	1	100	36	100	1125

$$\bar{d} = \frac{\sum d}{n} = \frac{-63}{10} = -6.3 \text{ and } S^2 = \frac{1}{n-1}\left[\sum d^2 - \frac{(\sum d)^2}{n}\right]$$

$$= \frac{1}{9}\left[1125 - \frac{(-63)^2}{10}\right] = 80.9$$

$$\therefore t = \frac{\bar{d}}{\sqrt{S^2/n}} = \frac{-6.3}{\sqrt{\frac{80.9}{10}}} = -2.214$$

$|t| = 2.214$ tabulated value $t_{0.05}$ for $t_{0.05}(10.1) = 9$ d.f, for two tail test is 2.26

Conclusion : Since $|t| < 2.26$, we accept H_0 and conclude that the training is not effective.

Example : 2.2. 2

A test was conducted to make changes in the typewriting machines. A group of typists were selected and were give two tests. One test was made on already used typewriting tables and another test was given using new typewriting machines. The difference in typed words per minute were registered

Typists	:	A	B	C	D	E	F	G	H	I
Increased No	:	2	4	0	3	-1	4	-3	2	5

Test whether changes made in the machines increased the typing speed.

Solution :

Null hypothesis : Ho : $\mu_1 = \mu_2$: (i.e) changes made in the type writing machine not increased the typing speed.

H_1 : $\mu_1 \neq \mu_2$ (two tailed)

PAIRED 't' TEST FOR SINGLE MEAN

Typists :	A	B	C	D	E	F	G	H	I	Total
d :	2	4	0	3	-1	4	-3	2	5	16
d^2 :	4	16	0	9	1	16	9	4	25	84

$$\bar{d} = \frac{\sum d}{n} = \frac{19}{9} = 1.778$$

$$S^2 = \frac{1}{n-1}\left[\sum d^2 - \frac{(\sum d)^2}{n}\right] = \frac{1}{8}\left[84 - \frac{(16)^2}{9}\right] = 6.9444$$

Test statistic

$$t = \frac{\bar{d}}{S/\sqrt{n}} = \frac{1.778 \times 3}{2.635} = 2.024$$

Tabulated value of $t_{0.05}$ for (9-1= 8d.f) is 2.306

Conclusion :

since $|t| < 2.306$, the null hypothesis is accepted and we conclude that changes made in typewriting machine has not increased the typing speed.

Exercise

1. Explain the procedure for testing the significant difference of paired t test for single mean
2. A certain stimulus administered to each of the 12 patients resulted in the following increase of blood pressure.

   5,2,8,-1,3,0,-2,1,5,0,4,6

can it be concluded that stimulus will, in general, be accompanied by an increase in blood pressure ?

(Ans $|t| = 2.89$, $t_{0.05,11} = 1.8$, reject H_0)

3. Twelve students were given intensive coaching and tests were given before and after the coaching. Is there any improvement due to coaching?

Marks in I test : 50 42 51 26 35 42 60 41 70 55 62 38 (before)

Marks in II test : 62 40 61 35 30 52 68 51 84 63 72 50 (after)

(Ans : $|t| = 4.885$, $t_{0.05,11} = 2.2$, reject H_0).

2.3 SIGNIFICANCE OF DIFFERENCE BETWEEN TWO SMALL SAMPLE MEANS

Suppose we want to test if two independent sample $x_1, x_2, \ldots x_{n_1}$ and $y_1, y_2 \ldots\ldots\ldots y_{n_2}$ have been drawn from two normal populations with means μ_1 and μ_1 respectively.

Under the null hypothesis H_0, the samples have been drawn from the normal populations having means μ_1 and μ_2 and their population variances are equal

(i.e $\sigma_1 = \sigma_2 = \sigma$) then test statistic

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $\bar{x}=\frac{\sum x}{n_1}, \bar{y}=\frac{\sum y}{n_2}$ and

$$S^2=\frac{1}{n_1+n_2-2}\left[\sum(x-\bar{x})^2+\sum(y-\bar{y})^2\right]$$

is an unbiased estimate of the population variance σ^2, then it follows t- distribution with n_1+n_2-2 degress of freedom.

If $|t|$ < table value of t_α, accept H_0 at α % level of significance.

Assumptions :

The following assumptions are made in using the test.

1. Parent populations, the samples drawn are normally distributed.
2. Population variances are equal and unknown.
3. The two samples are random and independent.

When sample SD's s_1 and s_2 are given, S^2 can be computed using the formula

$$S^2=\frac{n_1s_1^2+n_2s_2^2}{n_1+n_2-2}$$

Therefore test statistic t

$$t=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{n_1s_1^2+n_2s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

where $\bar{x}_1, \bar{x}_2$ are the means of two samples drawn from the normal populations.

Note :If $n_1 = n_2 = n$, then test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n-1}}}$$

Example : 2.3.1

Two independent sample of 8 and 7 items respectively had the following values

Sample I	9	11	13	11	15	9	12	14
Sample II	10	12	10	14	9	8	10	

Is the difference between the means of samples significant?

Solution :

Let x and y represent that variables of the samples I and II respectively.

Null hypothesis : $H_0 : \mu_1 = \mu_2$, there is no significance difference between two sample values

Alternative Hypothesis H_1: $:\mu_1 \neq \mu_2$ (two tailed test)

x	y	$(x-\bar{x})$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$
9	10	-2.75	7.5625	-0.428	0.1831
11	12	-0.75	0.5625	1.572	2.4711
13	10	1.25	1.5625	-0.428	0.1831
11	14	-0.75	0.5625	3.572	12.7591
15	9	3.25	10.5625	4.572	20.9031
9	8	-2.75	7.5625	-2.428	5.8951
12	10	0.25	0.0625	-0.428	0.1831
14					
94	73		28.4375		42.5777

$$\bar{x} = \frac{\sum x}{n_1} = \frac{94}{8} = 11.75 \quad \text{and} \quad \bar{y} = \frac{\sum y}{n_2} = \frac{73}{7} = 10.428$$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum (x - \bar{x})^2 + \sum y - \bar{y})^2\right]$$

$$= \frac{1}{13}[28.4375 + 42.5777] = 5.4627$$

Test statistic

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{11.75 - 10.428}{\sqrt{5.4627\left(\frac{1}{8} + \frac{1}{7}\right)}}$$

$$= \frac{1.322}{1.2095} = 1.0930$$

Tabulated value of t at 5% level of significance for 13 d.f is 2.16

Conclusion : Since $|t| < 2.16$, H_0 is accepted and we conclude there is no significant difference between the two sample means.

Example : 2.3.2

A reading rest is given to an elementary school class that consists of 12 Anglo- Indian children and 10 American – Indian children. The result of the test are

Anglo – Indian	American – Indian
$\bar{x}_1 = 74$	$\bar{x}_2 = 70$
$s_1 = 8$	$s_2 = 10$

Is the difference between the means of two groups significant at 5% level, given $t_{20} = 2.086$ $t_{22} = 2.074$ at 5% level

Solution :

Null Hypothesis : $H_0 : \mu_1 = \mu_2$ (i.e)

There is no significant difference between two samples means.

Alternative Hypothesis $H_1 : \mu_1 \neq \mu_2$

Test statistic : under H_0, the test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

where $S^2 = \frac{n_1 s_1^2 + n_1 s_2^2}{n_1 + n_2 - 2} = \frac{12 \times 64 + 10 \times 100}{12 + 10 - 2} = 88.4$

Therefore $t = \frac{74 - 70}{\sqrt{88.4\left(\frac{1}{12} + \frac{1}{10}\right)}} = \frac{4}{4.0253} = 0.9937$

Tabulated value of t at 13 d. f for 5% level of significance is 2.16

Conclusion:

Since $|t| < 2.16$, H_0 is accepted, and we conclude that is no significance

difference between two sample means.

Example : : 2.3.3

A Group of 10 rats fed on diet A and another group of 8 rats fed on diet B recorded the following increase in weight (gms)

Diet A :	5	6	8	1	12	4	3	9	6	10
Diet B :	2	3	6	8	10	1	2	8		

Does it show superiority of diet A over diet B ?

Solution :

Null Hypothesis : $H_0 : \mu_1 = \mu_2$ Assume that the difference in diets A and B is not significant

Alternative Hypothesis

$H_1: \mu_1 > \mu_2$ (single tailed test)

x_1	x_1^2	x_2	x_2^2
5	25	2	4
6	36	3	9
8	64	6	36
1	1	8	64
12	144	10	100
4	16	1	1
3	9	2	4
9	81	8	64
6	36		
10	100		
64	512	40	282

$$\bar{x}_1 = \frac{\sum x_1}{n_1} = \frac{64}{10} = 6.4$$

$$\bar{x}_2 = \frac{\sum x_2}{n_2} = \frac{40}{8} = 5$$

$$s_1^2 = \frac{1}{n_1}\sum x_1^2 - \bar{x}_1^2$$

$$= \frac{512}{10} - (6.4)^2 = 10.24$$

$$s_2^2 = \frac{1}{n_2}\sum x_2^2 - \bar{x}_2^2$$

$$= \frac{282}{8} - 25 = 10.25$$

$$S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \frac{10 \times 10.24 + 8 \times 10.25}{10 + 8 - 2} = 11.525$$

Test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{6.4 - 5}{\sqrt{11.525\left(\frac{1}{10} + \frac{1}{8}\right)}} = \frac{1.4}{1.6103} = 0.8694$$

Tabulated value of t at 10+8–2 = 16 d.f at 5% level of significance for single tailed test is 1.75 Since $|t| < 1.75$, we accept H_0 and we conclude that there is no significant difference between the average weights on diets A and B

Exercise

1. Explain the t-test for significance of the difference between the two sample means.
2. Samples of the two types of electric light bulbs were tested for length of life and following data were obtained

	Type -I	Type –II
Sample No:	$n_1 = 8$	$n_2 = 7$
Sample Mean	$\bar{x}_1 = 1{,}234 hrs$	$\bar{x}_2 = 1{,}036 hrs$

Sample SD's $s_1 = 36$ hrs $s_2 = 40$ hrs

Is the difference in the means significant to warrant that typer 11 is superior to type I regarding the length of life ?

(Ans: $|t| = 9.39$, H_0 is rejected)

3. Samples of two types of batteries were tested for length of life and following data were obtained.

	Type -I	Type –II
Sample No:	$n_1 = 9$	$n_2 = 8$
Sample Mean	$\bar{x}_1 = 600 hrs$	$\bar{x}_2 = 640 hrs$
Sample variances	$s_1^2 = 121$	$s_2^2 - 144$

It the difference between two means significant ?

(Ans : $|t| = 6.735$, not significant)

4. The following table gives the value of protein from Kangayam cow's milk and buffalo's milk Examine if the differences are significant.

Cow's milk :	1.90	1.95	2.00	2.02	1.85	1.80
Buffalo's milk :	2.12	2.00	2.20	2.45	2.20	2.10

(Ans $|t| = 3.6393$, the difference is significant)

5. A group of 5 patients treated with medicine A weigh 42, 39,48,60 and 41 kg. A second group of 7 patients from the same hospital treated with medicine B weight 38, 42,56,64,68,69 and 62 kgs. Do you agree with the claim that medicine B increase the weight significantly?

(Ans : $|t| = 1.70$ do not differ significantly at 5% level ofsignificance)

6. Below are given the gain in weights (in lbs) of pigs fed on two diets A and B Gain in weight

Diet A : 25, 32,30,34,24, 14,32,24,30,31,35,25

Diet B : 43, 34,22,10,47, 31,40,30,32,35,18,21,35,29,22

Test, is the two diets differ significantly as regards as their effect on increase in weight. (An. $|t| = 0.609$, two diets do not differ significantly as regards their effect on increase in weight)

1.4 SNEDCEOR'S F-TEST OF SIGNIFICANCE (FOR SMALL SAMPLES)

This test is used test the equality of the population variances.

If s_1^2 and s_2^2 are the variances of the two samples of sizes n_1 and n_2 respectively, the estimates of the population variance based on these samples are respectively

$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1}$ and $S_2^2 = \frac{n_2 s_2^2}{n_2 - 1}$. The quantities $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ are called the degrees of freedom of these estimates. We want to test if these estimates S_1^2 and S_2^2 are significantly different or if the samples may be regarded as drawn from the same population or from two populations with same variance σ^2.

Null Hypothesis :

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$ (i.e) population variances are same.

Computation of F- statistic

Under H_0, F - statistic

$$F = \frac{S_1^2}{S_2^2} \text{ here}$$

$$S_1^2 = \frac{1}{n_1 - 1}\sum (x - \bar{x})^2 = \frac{n_1 s_1^2}{n_1 - 1}$$

$$S_2^2 = \frac{1}{n_2 - 1}\sum (y - \bar{y})^2 = \frac{n_2 s_2^2}{n_2 - 1}$$

In the F-statistic, variance having greatest degrees of freedom is taken in the numerator. If tabulated value of F at ($v_1 = n_1 - 1$, $v_2 = n_2 - 1$) degrees of freedom at α level of significance is greater than the calculated value of F, null hypothesis H_0 will be accepted.

Example : 2.4.1

Two random samples from two different populations were taken.

Sample 1: 20 16 26 27 22 23 18 24 19 25

Sample II:27 33 42 35 32 34 38 28 41 43 30 37

Test whether the population variances are the same at 5% level of significance.

Solution :

Null Hypothesis :

Ho : $\sigma_1^2 = \sigma_2^2$ (ie) the two populations have the same variance

Alternative Hypothesis :

$H_1 : \sigma_1^2 \neq \sigma_2^2$ (two tailed)

$$\bar{x}_1 = \frac{\sum x_1}{n_1} = \frac{220}{10} = 22$$

$$\bar{x}_2 = \frac{\sum x_2}{n_2} = \frac{420}{12} = 35$$

x_1	$x_1 - \bar{x}_1$	$(x_1 - \bar{x}_1)^2$	x_2	$x_2 - \bar{x}_2$	$(x_2 - \bar{x}_2)^2$
20	-2	4	27	-8	64
16	-6	36	33	-2	4
26	4	16	42	7	49
27	5	25	35	0	0
22	0	0	32	-3	9
23	1	1	34	-1	1
18	-4	16	38	3	9
24	2	4	28	-7	49
19	-3	9	41	6	36
25	3	9	43	8	64
			30	-5	25
			37	2	4
220	0	120	420	0	314

Confidence level

$\alpha = 5\%$

Computation of F statistic

$$F = \frac{S_1^2}{S_2^2}$$

$$S_1^2 = \frac{1}{n_1 - 1}\sum (x_1 - \bar{x}_1)^2 = \frac{120}{9} = 13.83$$

$$S_2^2 = \frac{1}{n_2 - 1}\sum (x_2 - \bar{x}_2)^2 = \frac{314}{11} = 28.54$$

$S_2^2 > S_1^2$ Therefore $F = \frac{S_2^2}{S_1^2}$

$$= \frac{28.54}{13.83} = 2.14$$

Tabulated value $F_{0.05}(11,9) = 3.10$

Since calculated F < F 0.05, null hypothesis H_0 is accepted and we conclude that the two samples have been taken from the populations have same variance.

Example: 2.4.2

The two random sample reveal the following data

Sample No	Size	Mean	Variance
I	8	9.6	1.2
II	11	16.5	2.5

Test whether the samples came from the same population

Solution :

A normal population has two parameters mean μ and variance σ^2.If we want to test whether the independent samples have been drawn from the same normal population, we have to test

i. the equality of population means and

ii the equality of population variances

Null Hypothesis H_0 : The two samples have been drawn from the same normal population

(ie.) $\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$

To test if $\mu_1 = \mu_2$, we use 't' test and to test if $\sigma_1^2 = \sigma_2^2$ we use F-test. But to use t test the assumption is $\sigma_1^2 = \sigma_2^2$. Therefore before using t- test, we must take sure that the two populations have the same variances.

i. To test $\sigma_1^2 = \sigma_2^2$ (F- test)

Here $n_1 = 8$. $\bar{x}_1 = 9.6$, $s_1^2 = 1.2$

$n_2 = 11$, $\bar{x}_2 = 16.5$, $s_2^2 = 2.5$

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{8 \times 1.2}{8-1} = \frac{9.6}{7} = 1.37$$

$$S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{11 \times 2.5}{11-1} = \frac{27.5}{10} = 2.75$$

Since $S_2^2 > S_1^2$, F- Statistic is

$$F = \frac{S_2^2}{S_1^2} = \frac{2.75}{1.37} = 2.007$$

Tabulated value of $F_{.0.05}$ (10,7) d.f = 3.64 since calculated value of $F < F_{0.05}$, the null hypothesis H_0 is accepted and we conclude that the population variance are equal.

ii. To test H_0: $\mu_1 = \mu_2$

Alternative Hypothesis H_1 : $\mu_1 \neq \mu_2$ (two tailed)

$$S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \frac{8 \times 1.2 + 11 \times 2.5}{8 + 11 - 2} = \frac{9.6 + 27.5}{17} = 2.18$$

Test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{9.6 - 16.5}{\sqrt{2.18\left(\frac{1}{8} + \frac{1}{11}\right)}} = \frac{-6.9}{0.686} = -10.05$$

$|t| = 10.05$

Tabulated value of $t_{0.05}(17)$ d.f $= 2.11$

Since $|t| > 2.11$, H_0 is rejected (i.e) there is a significant difference in the population means. Therefore we conclude that the two samples have not been drawn from the same normal population.

Example : 2.4.3

In a sample of 8 observations, the sum of squares of deviations of the sample values from the sample mean was 94.5 and in another sample of 10 observations, it was found to be 101.7. Test whether the difference of variance is significant.

Null Hypothesis :

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$ (i.e) the estimates of σ^2 given by the samples are homogeneous $\sum (x_1 - \bar{x}_1)^2 = 94.5$

Therefore $S_1^2 = \frac{\sum (x_i - \bar{x})^2}{n_1 - 1} = \frac{94.5}{7}$

$\sum (x_2 - \bar{x}_2)^2 = 101.7$ therefore $S_2^2 = \frac{\sum (x_2 - \bar{x}_2)^2}{n_2 - 1}$

$$= \frac{101.7}{9} = 11.3$$

$$F = \frac{S_1^2}{S_2^2} = \frac{13.5}{11.3} = 1.195$$

Tabulated value of $F_{7,9}(0.05) = 3.29$

Since calculated value of F < tabulated value, we accept null hypothesis H_0 and we conclude that the difference is not significant.

Exercise

1. Two independent samples of sizes 9 and 7 from a normal population had the following values of the variables.

   Sample I : 18 13 12 15 12 14 16 14 15

   Sample II: 16 19 13 16 18 13 15

   Do the estimates of the population variance differ significantly at 5% level?

2. Two random sample reveal the following data.

Sample No.	Size	Mean	Variance
I	16	440	40
II	25	460	42

3. Write short notes on

   (i) F – test (ii) applications of F -test

4. Two random samples gave the following results

   $n_1 = 10 \quad \sum(x_1 - \bar{x}_1)^2 = 90 \quad \bar{x}_1 = 15, \ \bar{x}_2 = 14$

   $n_2 = 12 \quad \sum(x_2 - \bar{x}_2)^2 = 108$

   Test whether the samples came from the same population.

5. The nicotine content (in milligrams) of two samples of

tobacco were found to be as follows

Sample A :24 27 26 21 25

Sample B :27 30 28 31 22 36

Can it be said that the two samples come from the same normal population?

Ans : $H_0 : \mu_1 = \mu_2$; $t = 1.9$, Not significant

$H_0 : \sigma_1^2 = \sigma_2^2$, $F = 4.08 < 6.26$ $[F_{0.05}(5,4)]$ Not significant

6. Two independent samples of 8 and 7 items respectively has the following value of the variables

Sample I : 9 11 13 11 15 9 12 14

Sample II :10 12 10 14 9 8 10

Do the estimates of population variance differ significantly?

2.5 χ^2 TEST

Definition :

χ^2 test is based on non parametric test. This test was introduced by Karl pearson in the year 1900. This test is used to find the significant difference between observed and expected frequency values. It is defined as follows

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Where O denotes the observed frequency values and E denotes the expected frequency values.

χ^2 distribution :

Square of the standard normal variate is chi-square variate with 1 d.f. (i.e) If X follows a standard normal variate with mean μ and variance σ^2 then. $\left(\frac{x-\mu}{\sigma}\right)^2$ is a chi-square variable with 1 degrees of freedom. It depends on the no. of degrees of freedom. Different distributions are obtained for different degrees of freedom.

Properties of χ^2 distribution :

1. Mean of χ^2 distribution is the number of degrees of freedom (n).
2. Variance of χ^2 distribution is 2n.
3. Median of the χ^2 distribution divides the curve into two equal areas of values 0.5 on each side.
4. Characteristic function of χ^2 distribution is $(1-2\,it)^{-n/2}$
5. Moment generating function of χ^2 distribution is $(1-2\,t)^{-n/2}$
6. The sum of independent chi-square variables in also a chi-square variable.
7. χ^2 distribution tends to normal as $n \to \infty$
8. If χ_1^2 and χ_2^2 are two independent χ^2 variates with n_1, n_2 as degrees of freedom then $\frac{\chi_1^2}{\chi_2^2}$ is a β_2 variate.

9. The number of independent variables is usually called the number of degrees of freedom denoted by n.
10. When number of degrees of freedom n>30, $\sqrt{2\chi^2}$ approaches normal curve. Mean of $\sqrt{2\chi^2}$ is $\sqrt{2n-1}$ and its S.D. is 1.

Conditions of validity of chi-square test

1. N, the total of frequency, should reasonably large say greater than 50.
2. No theoretical cell frequency should be less than 5. If any theoretical cell frequency is less than 5, then for application of χ^2 test, it is pooled with the succeeding or preceding so that the pooled frequency is more than 5. This is to make the χ^2 distribution a continuous distribution to maintain the character of continuity of the distribution.
3. The sample observation should be independent
4. Constraints on the cell frequencies, if any, must be linear (eg $\sum O_i = \sum E_i$)

Application (or uses) of chi-square distribution

This distribution has large number of applications in statistics and some of them are listed below.

1. To test the goodness of fit
2. To rest the independence of attributes.
3. To test if the hypothetical value of the population variance is σ^2.

4. To test the homogeneity of independent estimates of the population variance.
5. To test the homogeneity of independent estimates of the population correlation co-efficient. We shall mainly use the first two applications.

Chi – square properties :

1. It is not symmetrical, it is skewed to the right.
2. Chi – square variate is non-negative in value; it is either zero or positively valued.

I.Goodness – of – fit – test

We shall consider the goodness – of –fit test as the first application of chi square. This test is a test of agreement (or conformity, or consistency) between a hypothetical and a sample distribution. A number of times we find that the results obtained in samples are not consistent with the theoretical results expected according to the rules of probability. For example, theory says that when we toss a fair coin 100 times, we should expect 50 heads and 50 tails, but in reality we may not find this result ; either heads or tails may be more than 50 and thus they may seldom equal. Before we take up an example, we give below procedure involved in the chi-square test.

Procedure for concluding the chi-square test

Step 1:

State the null hypothesis H_0 as the sample distribution agrees with

the hypothetical (or) theoretical distribution.

Step 2 :

Calculate the number in each category on the assumption that the null hypothesis is correct. Thus for each observation, we shall have observed frequency and expected frequency

Step 3 : determine the level of significance

Step 4: calculate the chi – squire by using the following formula

$$\chi^2 \quad \frac{\sum_{i=1}^{n}(O_i - E_i)^2}{E_i}$$

Where

O_i = observed frequency in the i^{th} category

E_i= Expected frequency in the i^{th} category

N = Number of categories

Step 5 :

Determine the number of degrees of freedom. For the specified level of significance and the degrees of freedom, find the critical or theoretical value of χ^2.

Step 6:

Compare the calculate value χ^2 with the theoretical value and determine the rejection. In case the calculated value of χ^2 is less than the theoretical value, the null hypothesis is accepted. If, calculated value of χ^2 is greater than the theoretical value, the null hypothesis H_0 is rejected.

Example : 2.5.1

Four coins were tossed at a time and then number of heads appeared were marked. The following data was obtained when the experiment was repeated 240 times.

Number of heads:	0	1	2	3	4
Frequencies:	13	64	85	58	20

Fit a Binomial distribution for this data and the unbiasedness of the coins.

Solution :

H_0 : The coins are unbiased.

We are given p = q = ½ / n = 4, and N = 240, To find expected frequencies

No of Heads	$P(X\text{-}x) = NC_x p^x q^{n-x}$ $P(x = x) = 4$	Expected frequency N.P (X = x)
0	$4C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^4 = \left(\frac{1}{16}\right)$	$\left(\frac{1}{16}\right) \times 240 = 15$
1	$4C_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^3 = \left(\frac{4}{16}\right)$	$\left(\frac{4}{16}\right) \times 240 = 60$
2	$4C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 = \left(\frac{6}{16}\right)$	$\left(\frac{6}{16}\right) \times 240 = 90$
3	$4C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^1 = \left(\frac{4}{16}\right)$	$\left(\frac{4}{16}\right) \times 240 = 60$
4	$4C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^0 = \left(\frac{1}{16}\right)$	$\left(\frac{1}{16}\right) \times 240 = 15$

Total = 240

Calculation for χ^2

Observed frequency O	Expected frequency E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
13	15	4	0.27
64	60	16	0.27
85	90	25	0.28
58	60	4	0.07
20	15	25	1.67

Therefore calculated value of X^2 : 2.56

Now tabulated value of chi square at (5-1= 4 d.f) is 9.488

Since calculated value of χ^2 is less than the tabulated values H_0 is accepted at 5% level of significance and we conclude that the coins are unbiased.

Example : 2.5.2

The following table shows the number of points obtained in a foot ball match

No of points	0	1	2	3	4	5	6	7
No of matches	95	158	108	63	40	9	5	2

Fit a poission distribution to the given data and test the goodness of fit.

Solution :

Null Hypothesis H_0 : The given data values fit the poission distribution.

Confidence level : Let $\alpha = 0.05$

Calculation of expected frequencies

$$m = \frac{\sum fx}{\sum f} = \frac{812}{480} = 1.7$$

In poisson distribution

$$P(X = x) = \frac{e^{-m} m^x}{x!} \quad x = 0,1,2,.....$$

$$\text{Now } P(0) = \frac{e^{-1.7}(1.7)^0}{0!} = 0.183$$

$f(0) = N.\ P(0) = 480 \times 0.184 = 87.84$

The other expected frequencies are obtained using recurrence relation

$$f(x+1) = \frac{m}{x+1} f(x)$$

when $x = 0$, $f(1) = \frac{m}{1} f(0) = 1.7 \times 87.85 = 149.328$

when $x = 1$, $f(2) = \frac{m}{2} f(1) = \frac{1.7 \times 149.328}{2} = 126.93$

when $x = 2$, $f(3) = \frac{m}{3} f(2) = \frac{1.7}{3} \times 126.93 = 71.927$

when $x = 3$, $f(4) = \frac{m}{4} f(3) = \frac{1.7}{4} \times 71.927 = 30.569$

when x = 4, $f(5) = \frac{m}{5} f(4) = \frac{1.7}{5} \times 30.569 = 10.393$

when x = 5, $f(6) = \frac{m}{6} f(5) = \frac{1.7}{6} \times 10.393 = 2.94$

when x = 6, $f(7) = \frac{m}{7} f(6) = \frac{1.7}{7} \times 2.94 = 0.719$

Observed frequency O	Expected frequency E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
95	88	49	0.56
158	150	64	0.43
108	126	324	2.57
63	72	81	1.13
40	30	100	3.33
9 5 16 2	10 3 14 1	4	0.29
			8.31

Chi-square test statistic

$$X^2 = \sum_{i=1} \frac{(O-E)^2}{E} = 8.31$$

Tabulated value of chi-square at 5% level of significance for (6-1-1=4 d.f) = 9.488. Since calculated χ^2 < tabulated value of χ^2, null hypothesis H_0 is accepted and we conclude that the given data values fits the Poission distribution.

Example : 2.5.4

A brand manager is concerned that her brand's share may be unevenly distributed throughout the country. In a survey, in which the country was divided into four geographical regions a random sampling of 100 consumers in each region was surveyed, with the following results :

Region

	NE	NW	SE	SW	Total
Purchase the brand	40	55	45	50	190
Do not purchase	60	45	55	50	210
Total	100	100	100	100	400

Develop a table of observed and expected frequencies for this problem.

1. Calculate the sample chi-square value
2. State the null and alternative hypothesis
3. At $\alpha = 0.05$, test whether branch share is the same across the four regions.

Solution :

The null Hypothesis H_0 : The brand share is evenly distributed.

Alternative Hypothesis H_1:The brand share is not evenly distributed.

The degree of freedom = (r-1)(c-1) = (2-1) (4-1) = 3

		O	E	$(O-E)^2$	$(O-E)^2/E$
Purchase	NE	40	47.5	56.25	1.18
the brand	NW	55	47.5	56.25	1.18
	SE	45	47.5	6.25	0.13
	SW	50	47.5	6.25	0.13
	NE	60	52.5	56.25	1.07
Do not	NW	45	52.5	56.25	1.07
purchase	SWE	55	52.5	6.25	0.12
	SW	50	52.5	6.25	0.12
					5.00

The critical value of χ^2 at $\alpha = 0.05$ level for 3 degrees of from the tables is 7.815

Since calculated χ^2 < tabulated χ^2, we accept null hypothesis H_0 and conclude that the brand share is evenly distributed in all the four regions of the country.

to the genetic model these numbers should be in the ratio 9:3:4. Are the data consistent with the model at 5% level?

2.6 : χ^2 TEST OF INDEPENDENCE OF ATTRIBUTES

In the test of independence of attributes the population and sample are classified according to several attributes. We are interested in ascertaining whether there is any dependency relationship between the two attributes. We apply chi-square distribution, it will indicate whether or not the two attributes are independent. If our chi-square test indicates independence, it means

the observed data are consistent with the hypothesis that the two attributes are independent.

Formula for computing chi-square(χ^2)

Here we introduce 2x2 table for calculating χ^2. The cell frequencies and marginal totals are shown in the following table

Observed frequency Table

a	b	(a+b)
c	d	(c+d)

(a+c) (b+d) N = a+b+c+d

Expected frequency table

$\frac{(a+c)(a+b)}{N}$	$\frac{(b+d)(a+b)}{N}$
$\frac{(a+c)(c+d)}{N}$	$\frac{(b+d)(c+d)}{N}$

Chi-square can be calculated by a simple formula, which involves only the observed frequencies.

$$\chi^2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

Example : 2.6.1

A lawn equipment shop is considering adding a brand of lawn movers to its merchandise. The manager of the shop believes that the highest quality lawn movers are trooper. Lawn Eater and Nipper and he needs to decide whether it makes a difference which of these three shop adds to its existing merchandise.

Twenty owners of each of these three types of lawn movers are randomly sampled and asked how satisfied they are with their lawn movers.

Lawn mover	Very satisfied	Satisfied	Not satisfied	Total
Trooper11	11	6	3	20
Lawn eater	13	4	3	20
Nipper	13	6	1	20

Are the owners of the lawn movers homogenous in their response of the survey? Use a 5% significance level.

H_0 : Owners of the lawn movers homogeneous in this response to the server.

H_1 : Owners of the lawn movers are not homogeneous in the response to the server.

We combine satisfied and not satisfied category as not satisfied category.

	Very satisfied	Not satisfied	
Transfer	11	9	20
Lawn Movers	13	7	20
Nipper	13	7	20
Total	37	23	60

O	E	$(O-E)^2$	$(O-E)^2/E$
11	12.33	1.7689	0.143
9	12.33	1.7689	0.143
13	12.33	1.7689	0.143
7	7.66	0.4356	0.056
13	7.66	28.515	3.722
7	7.66	0.4356	0.056
		χ^2	4.263

Calculated value of χ^2 at (3-1)(2-1) = 2x1 = 2 d.f at 5% level of significance is 5.991

Since calculated value of χ^2 < tabulated value we accept null hypothesis H_0 and we conclude that owners of lawn movers are homogeneous in this response to the server.

Example : 2.6.2

On the basis of information noted below, find out whether the new treatment is comparatively superior to the conventional one

	Favourable	**Non favourable**	**Total**
Conventional	40	70	110
New	60	30	90
Total	100	100	200

Solution :

Null Hypothesis : H_0 : The two treatments are independent.

$$^2 = \frac{(ad-bc)^2 \times N}{(a+c)(b+d)(a+b)(c+d)}$$

$$\chi^2 = \frac{(40\times30-60\times70)^2(200)}{(100)(100)(110)(90)} = 18.1818$$

For a 2x2 contingency table, d.f v = (r-1) (c-1) = (2-1) (2-1) = 1, χ^2 tabulated value at 5% level of significance for 1 d.f = 3.81.4 Since calculated value of χ^2 > tabulated value of χ^2, we reject H_0 and we conclude that, the treatments are not independent.

Example : 2.6.3

From the following data, find whether hair colour and sex are associated.

	Hair colour				
	Fair	**Red**	**Medium**	**Dark**	**Black**
Boys	592	119	849	504	36
Girls	544	97	677	451	14

Solution :

Null Hypothesis H_0 : Hair colour and sex are not associated.

	Fair	**Red**	**Medium**	**Dark**	**Black**	**Total**
Boys	592	119	849	504	36	2100
Girls	544	97	677	451	14	1783
Total	1136	216	1526	955	50	3883

Expected frequency values are

$$E(592) = \frac{1136\times2100}{3883} = 614.3$$

E(119) = 116.8 E (849) = 825.2 E (504) = 516.4 E(36) = 27.04
E (544) = 521.6 E (97) = 99.2 E(677) = 700.7 E (451) = 438.5
E (14) = 23

O	E	$(O-E)^2$	$(O-E)^2/E$
592	614	484	0.788
119	117	4	0.034
849	825	576	0.698
504	516	144	0.2719
36	27	81	3.000
544	522	484	0.927
97	99	4	0.040
677	701	576	0.821
451	439	144	0.328
14	23	81	3.521
			$\chi^2 = 10.436$

Degrees of freedom = (2-1)(5-1) = 1 x4 = 4 d.f

Tabulated $\chi^2_{0.05}$ (4) = 9.488

Since calculated χ^2 > tabulated χ^2, H_0 is rejected at 5% level of significance and we conclude that hair colour and sex are not associated.

Example : 2.6.4

1072 college students were classified according to their intelligence and economic conditions. Test whether there is any

χ^2 TEST OF INDEPENDENCE OF ATTRIBUTES

		Intelligence				
		Excellent	Good	Medicore	Dull	Total
Economic Conditions	Good	48	199	181	82	510
	Not Good	81	185	190	106	562
	Total	129	384	371	188	1072

Solution :

Null Hypothesis H_0 : Intelligence and economic conditions are independent.

Expected frequencies are

E(48) = 61 E(199) = 183 E (181) = 177 E(82) = 89

E(81) = 68 E(185) = 201 E (190) = 194 E(106) = 99

O	E	$(O-E)^2$	$(O-E)^2/E$
48	61	169	2.770
199	183	256	1.398
181	177	16	0.090
82	89	49	0.550
81	68	169	2.485
185	201	256	1.273
190	194	16	0.082
106	99	49	0.494
		χ^2	9.142

Calculated $\chi^2 = 9.142$

Degrees of freedom = (2-1) (4-1) = 3

Tabulated $\chi^2_{0.05}$ (3) d.f = 7.815

Conclusion :

Since calculated χ^2 > tabulated χ^2 H_0 is rejected and we conclude that intelligence and economic conditions are not independent.

Example : 2.6.5

For the data in the following table test for independence between a person's ability in mathematics and interest in economics.

Ability in mathematics

		Low	Average	High	Total
Interest in	Low	63	42	15	120
Economics	Average	58	61	31	150
	High	14	47	29	90
	Total	135	150	75	360

State clearly the assumptions underlying your test procedure.

Solution :

Null Hypothesis H_0: Ability in Mathematics and interest in Economics are independent.

E(163) = 45 E(42) = 50 E (15) = 25 E(58) = 56

E(61) = 63 E(31) =31 E(14) = 34 E(47) = 38

E(29) = 19

O	E	$(O-E)^2$	$(O-E)^2/E$
63	45	324	7.2
42	50	64	1.28
15	25	100	4
58	56	4	0.071
61	63	4	0.063
31	31	0	0
14	33.75	390.0625	11.557
47	37.5	90.25	2.406
29	18.75	105.0625	5.603

$\chi^2 = 32.18$

Calculated value of $\chi^2 = 32.18$

Degrees of freedom $= (3-1)(3-1) = 4$

Calculated value of $\chi^2_{0.05}$ (4) d.f = 9.488

Since calculated value of χ^2 is > tabulated χ^2, null hypotheses H_0 is rejected and is concluded that ability in Mathematics and interest in Economics are not independent.

Exercise

1. The table below gives the number of aircraft accidents that occurred during the various days of the week. Test whether the accidents are uniformly distributed over the week.

Days	: Mon	Tues	Wed	Thurs	Fri	Sat
Accidents:	14	18	12	11	15	14

2. In an experiment on information of cattle from tuberculosis the following results were obtained.

	Affected	Unaffected	Inoculated
		12	28
Not inoculated		13	7

Examine the effect of variance in controlling the incidence of the disease.

3. The theory predicts the proportion of beans in the four groups A,B,C,D should be 9:3:3:1 in an experiment among 1600 beans, the numbers in the four groups were 882, 313,287 and 118. Does the experimental result support the theory result.

4. A survey of 320 facilities with 5 children each recorded the following distribution.

No of boys	: 5	4	3	2	1	0
No of girls	: 0	1	2	3	4	5
No of families:	14	56	110	88	40	12

Is this result consistent with the hypothesis that male and female births are equally probable?

5. Fit a Poisson distribution to the following data and test the goodness of fit.

X :	0	1	2	3	4	5	6
f:	275	72	30	7	5	2	1

6. The following table gives for a sample for married women, the level of education and marriage adjustment score.

		Very low	Low	High	Very high
Level of Education	College	24	97	62	58
	High School	22	28	30	41
	Middle School	32	10	11	20

Can you conclude from above, the higher the level of education, the greater is the degree of adjustment in marriage?

7. In an investigation into the health and nutrition of two groups of children of different social status, the following results are got.

Social status / Health	Poor	Rich	Total
Below normal	130	20	150
Normal	102	108	210
Above Normal	24	96	120
Total	256	224	480

Discuss the relation between the health and their social status.

8. The following table gives a classification of a sample of 160 plants of their flower colour and flatness of leaf.

	Flat Leaves	Curled leaves	Total
White flower	99	36	135
Red Flower	20	5	25
Total	119	41	160

Test whether the flower colours is independent of the flatness of leaf.

9. Two researchers A and B adopted different techniques while rating the students level. Can you say that the techniques adopted by them are significant?

Researchers	Below average	Average	Above average	Genius	Total
A	40	33	25	2	100
B	86	60	44	10	200
Total	126	93	69	12	300

10. A certain drug was administered to 500 people out of a total of 800 included in the sample to test its efficacy against typhoid the results are given below:-

	Typhoid	No typhoid	Total
Drug	200	300	500
No Drug	280	20	300
Total	480	320	800

On the basis of this data, can we say that drug is effective in preventing typhoid?

11. 200 digits were chosen at random from a set of tables. the frequencies of digits are

Digit	:	0	1	2	3	4	5	6	7	8	9
Frequency:		18	19	23	21	16	25	22	20	21	15

Use χ^2 test to assess the correctness of the hypothesis that the digits were distributed in equal numbers in the tables from which these numbers were taken

(Ans : $\chi^2 = 4.3$, H_0 rejected)

CHAPTER III

SAMPLING TECHNIQUE

3.1 Introduction

In real life situations the sampling technique is essential to face the challenges before discussing various techniques we list some definitions and basic concepts.

Definition :

Population (or universe) is the aggregate of statistical data forming a subject of estimation

A **sample** is a group of certain items taken from a large population in a specified manner. It is a portion of a population which is examined with a view to estimating the characteristics of the population.

There are two ways of collection of statistical data, viz, by complete enumeration (or) census and sample survey.

When information is collected in respect of every individual item, the enquiry is said to be done by complete enumeration (or) census. Eg. Estimating population 10 years once. But this method involved huge amount of money, large time, and much effort. In most cases of statistical inquiry, because of limitation of time and cost, only portion (ie a sample) of the available source of information is examined and data collected

from this process of partial enumeration is known as sample survey.

The results and findings are generalized and made applicable to the whole field of enquiry. This is known as sampling.

The process of inferring something about large groups of elements by studying only a part of it is known as **sampling.**

The process of sampling is based on two important principles of statistics.

i. Law of statistical regularity.

ii. Law of inertia of large number.

Sampling Distribution

Consider all possible samples of size n which can be drawn from a given sample .For each sample we can compute a statistic, such as mean, standard deviation etc which will vary from sample to sample. The aggregate of various values of the statistic under consideration may be grouped into a frequency distribution. This distribution is known as **Sampling distribution** of the statistic. Thus the probability distribution of all possible values that a sample statistic, is called the sampling distribution of the statistic. Sample mean and sample proportion based on a random samples are examples of sample statistic.

Suppose a Market Research Agency wants to estimate annual household expenditure on consumer durables from among the population of house holders.(say 25000) households in

TamilNadu. The agency can choose 10 different samples of 50 households each. For each of the samples, we can calculate the mean expenditure on consumer durables as given in the following table.

Sample No.	Total Expenditure for 50 households	Mean (Rs.)
1	10000	2000
2	30000	6000
3	20000	4000
4	15000	3000
.	.	.
.	.	.
.	.	.
49	60000	12000
50	8000	8000

The distribution of all sample means is known as the **sampling distribution of the mean.**

Similarly, the sampling distribution of the sample variance and the sample proportion can also be obtained.

Methods of Sampling

Sampling

↓

Probability Sampling (A particular element is included in the sample or not is governed by chance alone)	Non probability sampling (samples may be picked up based on judgment (or) convenience of Enumerator)
↓	↓
1. simple random sampling	1. Convenience sampling
2. systematic sampling	2. Purposive sampling
3. stratified sampling	3. Judgment sampling
4. cluster sampling	4. Sequential sampling

Probability sampling

3.2. SIMPLE RANDOM SAMPLING

In this process a sample of size 'n' is chosen and equal probability of being picked up as the chosen sample. To illustrate this idea we take a family of 6 members consisting of 2 adults and 4 children viz, A,B,C,D,E,F respectively. Total No. of ways to pick up,5 members of the family = $6C_5 = 6$ ways. And probability of each 5

members can be selected is $\frac{1}{6}$

$$P(ABCDE) = \frac{1}{6}$$

$$P(ABCDF) = \frac{1}{6}$$

$$P(ACDEF) = \frac{1}{6}$$

$$P(BCDEF) = \frac{1}{6}$$

$$P(ABCEF) = \frac{1}{6}$$

Probability of element A to be in the sample

$= P(A) = P(ABCDE) + P(ABCDF) + P(ACDEF) + P(ABDEF) +$

$P(ABCEF) = \frac{5}{6}$

Similarly, $P(B) = \frac{5}{6}$, $P(C) = \frac{5}{6}$, $P(D)$

$= \frac{5}{6}$, $P(E) = \frac{5}{6}$ and $P(F) = \frac{5}{6}$.

In general, if N = total number of elements in a population and n = no of elements to be chosen out of it, then the probability that particular element (say A) included in the sample

$$\frac{(N-1)C_{n-1}}{NC_n} = \frac{n}{N}.$$

In fact that every element of the population has an equal probability of getting included in the sample is made of use, in the simple random sapling technique.

3.3. SYSTEMATIC SAMPLING

Systematic sampling proceeds selecting one element after a fixed, interval depending on the sample ratio. For, example, if we want to have a sample of size 5 from a population of size 100, our sampling ratio would be $\frac{5}{100} = \frac{1}{20}$. We would therefore, have to decide where to start from among first 20 names from our frame. If this number happens to be 5 for example, then the sample would contain members having serial numbers5, 25,45,65,85 in the frame. Here the random process establishes only the first member of the sample the rest are pre – ordinated automatic because of known sampling ratio.

3.4 STRATIFIED RANDOM SAMPLING

It is generally used when the population is heterogeneous. Various steps involved in the stratified random sampling are as follows.

1. The population first sub divided into several parts (or small groups) called stratas according to some relevant characteristics so that each stratum (sub-population) is more or less homogeneous.
2. A small group (called sub-sample) is selected from each striatum at random.
3. All the sub samples combined together form the stratified sample.

Use. 1. it represent the population properly.

Definition :

The process of obtaining and estimating a stratified sample with a view to estimating the characteristic of the population is known as **stratified sampling.**

For example, for a stratified sample of 100 families from a city with a view to studying their economic condition. The city area is divided into a number of strata according to the annual income of the inhabitants. Thus localities mostly inhabited by people with more or less similar annual income may be included under one stratum.

Parameter and statistic

Statistical constants of the population viz. mean (μ) variance (σ^2) etc are referred as parameter. Statistical measures computed from the sample observations alone (eg) mean ($\bar{x}$) variance (s^2) etc are called sample statistic by professor R.A fisher.

In practice parameter values are not known and estimates based on sample values are generally used. Thus statistic which may be reached as an estimate of parameter, obtained from the sample, is a function of the sample values only.

3.5 STANDARD ERROR

The standard deviation of the sampling distribution of a statistic is called standard error of the statistic.

The standard deviation of the sampling distribution of the sample mean is called the standard error of the mean.

The standard deviation of the distribution of the sample proportions is called the standard error of the proportion.

Standard error is popularly known as sampling error.

Remark :

The standard error is inversely proportional to the sample size (ie) larger the sample size, the smaller the standard error.

Uses of Standard Error

1. Sampling error throws light on the precision and accuracy of the estimate.
2. It measures the dispersion of possible values of the statistic. in repeated samples of a fixed size from a given population.
3. It is used to set up confidence limits for population parameter. in tests of significance.

For example, S.E of the sample mean $\bar{x}$ and sample proportion p. are used to find confidence limits for the population mean μ and population proportion P.

S.No	Statistic	Standard Error
1.	Sample mean : $\bar{x}$	$\sigma/\sqrt{n}$
2.	Observed sample proportion : p	$\sqrt{PQ/n}$

3	Sample s.d :s	$\sqrt{\sigma^2/2n}$
4	Sample variance :	$\sigma^2/2n$
5	Difference of two sample means : $(\bar{x}_1 - \bar{x}_2)$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
6	Difference of two sample s.d's (s_1-s_2)	$\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$
7	Difference of two sample proportions(p_1-p_2)	$\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$

3.6 : SAMPLING WITH AND WITHOUT REPLACEMENT

If an element is picked up once and if it is not available to be picked up again, it is a sampling without replacement.

In the sampling with replacement, an element after being picked up and included in the sample, is replaced in the population so that it can be picked up again. The methods of selecting a random sample are i. Lottery method ii. Use of table of random numbers (see appendix).

Statistic	Standard Error	Remarks
Sample Mean $\bar{x}$	$\frac{\sigma}{\sqrt{n}}$	Population size is infinite (or) sample with replacement

	$\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$	Population size N finite (or) sample without replacement
Sample proportion p	$\sqrt{\frac{PQ}{n}}$	Population size is infinite (or) sample with replacement.
	$\sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}}$	Population size N finite (or) sample without replacement.

Example : 3.6.1

A population consists of the 5 numbers 3,7,11 and 15. Consider all possible samples of size 2, which can be drawn with replacement from this population. Find

i. the population mean

ii. the population standard deviation.

iii. the mean of the sampling distribution of the mean.

iv.the standard deviation of the sampling distribution of the mean.

Solution:

i. The population mean $\mu = \frac{\Sigma x}{n} = \frac{3+7+11+15}{4} = 9$

ii. The variance of the population $\sigma^2 = \frac{1}{N}\Sigma(x-\mu)^2$

$$= \frac{1}{4}\{(3-9)^2 + (7-9)^2 + (11-9)^2 + (15-9)^2\} = 20$$

Therefore the standard deviation of the population $\sigma = 4.472$

iii. There are 16 samples of size two which can be drawn with replacement. They are be drawn with replacement. They are

(3,3)	(3,7)	(3,11)	(3,15)
(7,3)	(7,7)	(7,11)	(7,15)
(11,3)	(11,7)	(11,11)	(11,15)
(15,3)	(15,7)	(15,11)	(15.15)

The corresponding sample means are

3.0	5.0	7.0	9.0
5.0	7.0	9.0	11.0
7.0	9.0	11.0	13.0
9.0	11.0	13.0	15.0

The mean of sampling distribution of means

$$\mu_{\bar{x}} = \frac{\text{Sum of all sample means}}{16}$$

$$= \frac{144}{16} = 9$$

iv. The variance of $\sigma_{\bar{x}}^2$ of the sampling distribution of means is

$$\sigma_{\bar{x}}^2 = \frac{1}{16}\{(3-9)^2 + (5-9)^2 + (7-9)^2 + \ldots + (13-9)^2 + (15-9)^2\}$$

$$\sigma_{\bar{x}}^2 = \frac{1}{16}[160] = 10$$

Therefore standard deviation of the sampling distribution of mean $\sigma_{\bar{x}}^2 = \sqrt{10} = 3.162$.

Example : 3.6.2

The weights of 1500 iron rods are normally distributed, with mean of 22.4 legs and standard division of 0.048 kgs. If 300 random samples of size 36 are drawn from this population , determine the mean and standard deviation of the sampling distribution of the mean. when sampling is done (i) with replacement (ii) without replacement.

Solution :

Given N = 1500, $\mu = 22.4$, $\sigma = 0.048$ n = 36

i. Sampling with replacement $\mu_{\bar{x}} = \mu = 22.4kg$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.048}{\sqrt{36}} = \frac{0.048}{6} = 0.008kg$$

ii. Sampling without replacement

The mean of the sampling distribution of the means is

$$\mu_{\bar{x}} = \mu = 22.4kg$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

$$= \frac{0.048}{\sqrt{36}} \times \sqrt{\frac{1500-36}{1500-1}}$$

$$= 0.008 \sqrt{\frac{1464}{1499}}$$

$$= 0.00796 \text{ kg}$$

Example : 3.6.3

2 % of out going Maths CA students in a college have joined M.Sc Statistics in I.S.I kolkatta. What is the probability that in a group of 700 such students 3% or more will be joining in I.S.I kolkatta.

Solution :

Here N is not given, but n = 700

population proportion P = 2 % = 0.02

Therefore Q = 1 - P – 0.08

The sample size is large

Therefore the sample proportion is normally distributed with mean

$\mu_p = 0.02$ and S.D. $\sqrt{\frac{PQ}{n}} = \sqrt{\frac{0.02 \times 0.08}{700}} = 0.00529$

Probability that sample proportion $p \geq 0.03$

= Area under the normal curve to the right of z = 1.89

= 0.5 – 0.4706

= 0.0294 $\quad \because z = \frac{p - P}{S.D} = \frac{0.03 - 0.02}{0.00529} = 1.89$

Diagram

Example : 3.6.4

The marks obtained by the Maths CA department students of a college in the paper statistics maths in the Bharathiar university semester examination are normally distributed with a mean of 75 and a standard deviation of 37.5 A random sample of 45 students is drawn from this population :

i) What is the standard error of the sampling mean?

ii) What is the probability that the mean of a sample of 9 students will be ether less than 57.5 or grater than 95?

Solution

i. The standard error of the sample mean $\bar{x}$ is given by

$$\sigma_{\bar{x}} = \frac{\sigma}{n} = \frac{37.5}{\sqrt{45}} = 5.59$$

ii. The random variable $\bar{x}$ follows normal distribution with mean $\mu_{\bar{x}}$ and standard deviation $\frac{\sigma}{\sqrt{n}}$

To find $P(\bar{x} < 57.5 \text{ or } \bar{x} > 95)$

$$= P(\bar{x} < 57.5) + P(\bar{x} > 95)$$

$$P\left(\frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} > \frac{57.5 - 75}{5.59}\right) + P\left(\frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} > \frac{95 - 75}{5.59}\right)$$

$$= P(z < 3.13) + P(z > 3.58)$$

$$= [(0.5 - P(0 < z < 3.13)] + 0.5 - P(0 < z < 3.58)$$

$$= (.5 - 0.4991) + (.5 - 0.4998)$$

$$= 0.0011$$

Example : 3.6.5

A random sample of size 7 is drawn without replacement from a finite population consisting of 88 units. If the population

S.D is 9.23, Find the S.E. of the sample mean.

Solution :

population size N = 88

sample size n = 7

standard deviation of the population = $\sigma = 9.23$

S.E. of the sample mean

$$= \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \qquad (\because \text{ N is finite})$$

$$= \frac{9.23}{\sqrt{7}} \sqrt{\frac{88-7}{88-1}} = 3.366$$

EXERSCISE

1. A population consists of 5 numbers 2,3,6,8,11. Consider all possible samples of size 2 each can be drawn with replacement from this population . find (i) mean of the population (ii) the standard deviation of the population (iii) the mean of the sampling distribution of means and (iv) the standard error of the means.

   Ans: $\mu = 6$, $\sigma^2 = 10.8$. $\sigma = 3.29$, $\mu_{\bar{x}} = 6.0$, $\sigma_{\bar{x}}^2 = 2.32$

2. Assume that monthly savings of 1000 employees working in a factory are normally distributed with mean Rs. 2000 and standard deviation Rs. 50. If. 25 sample consisting of 4 employees each are obtained, what would be the mean and standard deviation of the resulting sampling distribution of

means if sampling were done (i) with replacement (ii) with out replacement

Ans:

(i). $\mu_{\bar{x}} = \mu = 2000, \sigma_{\bar{x}} = 25$

(ii). $\mu_{\bar{x}} = \mu = 2000, \sigma_{\bar{x}} = 24.96$

3) 1% of the screws produced by a machine or defective. what is the probability that in a consignment of 500 such screws, 2% or more will be defective.

Ans : 0.0062

4) The marks obtained by students in an aptitude test are normally distributed with a mean of 60 and standard deviation of 30. Random sample of 36 students is drawn from this population

i) what is the standard error of the sampling mean ?

ii) what is the probability that the mean of a sample of 16 students will be wither les than 50 or greater than 80? $P(0 < Z < 4 = 0.4999)$

Ans : (i). $\sigma_{\bar{x}} = 5$

(ii). $P[\bar{x} < 50 \text{ or } \bar{x} > 80) = 0.02283$

5) A random sample of size 5 is drawn without replacement from a finite population consisting of 41 units. If the population S.D is 6.25 find the S.E of the sample mean.

$$\text{Ans : S.E} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{6.25}{\sqrt{5}}\sqrt{\frac{41-5}{41-1}} = 2.65$$

3.7 THEOREMS ON SIMPLE RANDOM SAMPLING

In this section we assume each unit of the population has an equal and independent chance of being included in the sample, also an equal probability of selection is assigned to each unit of the population at first drawn. It also implies an equal probability of selecting any unit from the available units at subsequent draws.

The probability of drawing any unit from the population of N units is 1/N.

The probability of drawing any unit in the second draw from among the available (N-1) units is $\frac{1}{N-1}$ and so on.

Let E_r be the even that any specified unit is selected at the r^{th} draw.

Therefore $P(E_r)$ = Prob (that the specified unit is not selected in anyone of the previous (r-1) draws and then selected in the r^{th} draw).

$P(E_r) = \prod_{i=1}^{r-1} P\{$ it is not selected in the i^{th} draw$\}$ P (It is selected at the r^{th} draw given that it is not selected at the previous r-1 draws)

By compound probability theorem, since draws are independent.

$$P(E_r) = \left(1-\frac{1}{N}\right)\left(1-\frac{1}{N-1}\right)\cdots\cdots\left(1-\frac{1}{N-(r-2)}\right)\times\frac{1}{N-(r-1)}$$

$$= \frac{(N-1)}{N}\cdot\frac{(N-2)}{(N-1)}\cdot\frac{(N-3)}{(N-2)}\cdots\cdots\frac{N-r+1}{N-r+2}\cdot\frac{1}{N-r+1} = \frac{1}{N}$$

Therefore $P(E_r) = \frac{1}{N} = P(E_1)$

This leads to a very interesting property.

The probability of selecting a specified unit of the population at any given draw is equal to the probability of its being selected at the first draw.

Notations and terminology :

Let us consider a finite population $U_1, U_2, \ldots\ldots\ldots U_n$ of size N and a sample $u_1, u_2 \ldots\ldots u_n$ of size n.

Let Y_i (i = 1,2,...N) be the value of the character for i^{th} unit in the population and y_i (i = 1,2...n) denote the value of the character for the i^{th} unit in the sample. Then define the population mean.

$$\bar{Y}_N = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

Sample Mean $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$

Sample Mean $\bar{y}_n$ may be alternatively written as

$$\bar{y}_n = \frac{1}{n}\sum_{i=1}^{N} a_i Y_i$$

Where a_i = { 1 if i^{th} unit is included in the sample

= 0 if i^{th} unit is not included in the sample

s^2 = Mean square of the population

$$= \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y}_N)^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} Y_i^2 - N\bar{Y}_N^2\right]$$

s^2 = Mean square of the sample

$$= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\bar{y}_n^2\right]$$

Theorem 3.7(i)

In sample random sampling without replacement (srswor) the sample mean is an unbiased estimate of the population mean.

(i.e) $E(\bar{y}_n) = \bar{Y}_N$

Proof :

$$E(\bar{y}_n) = E\left[\frac{1}{n}\sum_{i=1}^{N} a_i Y_i\right]$$

$$= \frac{1}{n}\left[\sum_{i=1}^{N} E(a_i) Y_i\right]$$

Since a_i takes only two values 1 and 0

$E(a_i)$ = 1.P (i^{th} unit is included in a sample of size n)

+ 0.P (i^{th} unit is not included in a sample of size n)

$$= 1\left(\frac{n}{N}\right) + 0\left(1 - \frac{n}{N}\right) = \frac{n}{N}$$

Hence $E(\bar{y}_n) = \left[\sum_{i=1}^{N} \frac{n}{N} . Y_i\right]\frac{1}{n}$

$$= \frac{1}{n} . \frac{n}{N}\sum_{i=1}^{N} Y_i$$

$$= \frac{1}{N} . N\bar{Y}_N$$

$$= \bar{Y}_N$$

Theorem :3.7(ii)

In srswor, the sample mean square is an unbiased estimate of the population mean square $E(s^2) = S^2$

Proof : $s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\bar{y}_n^2\right]$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i^2 + \sum_{i\neq j=1}^{n} y_i y_j\right)\right]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} y_i^2 -$$

$$= \frac{1}{n(n-1)}\sum_{i=1}^{n} y_i^2 - \frac{1}{n(n-1)}\sum_{i\neq j=1}^{n} y_i y_j$$

$$= \frac{(n-1)}{n}\cdot\frac{1}{(n-1)}\sum_{i=1}^{n} y_i^2 - \frac{1}{n(n-1)}\cdot\sum_{i\neq j=1}^{n} y_i y_j$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \frac{1}{n(n-1)}\sum_{i\neq j=1}^{n} y_i y_j \quad \ldots\ldots(1)$$

$$E(s^2) = \frac{1}{n}E\left(\sum_{i=1}^{n} y_i^2\right) - \frac{1}{n(n-1)}E\left[\sum_{i\neq j=1}^{n} y_i y_j\right] \ldots\ldots(2)$$

$$\text{Now } E\left[\sum_{i=1}^{n} y_i^2\right] = E\left[\sum_{i=1}^{N} a_i Y_i^2\right] = \sum_{i=1}^{N} E(a_i) Y_i^2 \quad \ldots\ldots (3)$$

a_i is defined as

a_i = 1, if i^{th} unit is included in the sample

0, if i^{th} unit is not included in the sample

$$\text{and } E\left[\sum_{i \neq j=1}^{n} y_i y_j\right] = E\left[\sum_{i \neq j=1}^{N} a_i a_j Y_i Y_j\right]$$

$$= \sum_{i \neq j=1}^{N} E(a_i a_j) Y_i Y_j$$

Now $E(a_i a_j) = 1.\ P(a_i a_j = 1) + 0.P(a_i a_j = 0)$

$= P(a_i = 1).\ P(a_j = 1 \mid a_i = 1) + 0$

$= P(Y_i$ is included in the sample of size n) $\times$

$P(Y_j$ is included in the sample given that Y_i is included in the sample)

$$E(a_i a_i) = \frac{n}{N} \cdot \frac{n-1}{N-1}$$

$$E\left[\sum_{i \neq j=1}^{n} y_i y_j\right] = \sum_{i \neq j=1}^{N} E(a_i a_j) Y_i Y_j$$

$$= \sum_{i \neq j=1}^{N} \frac{n(n-1)}{N(N-1)} Y_i Y_j$$

$$= \frac{n(n-1)}{N(N-1)} \sum_{i \neq j=1}^{N} Y_i Y_j \qquad \ldots(4)$$

$$= E(a_i) = 1.\ \frac{n}{N} + 0.\left(1 - \frac{n}{N}\right)$$

Substitute in (3) $E(a_i) = \frac{n}{N}$

Therefore $E\left[\sum_{i=1}^{n} y_i^2\right] = \sum_{i=1}^{N} \frac{n}{N} Y_i^2$

$$= \frac{n}{N} \sum_{i=1}^{N} Y_i^2 \qquad \ldots.(5)$$

Sub (4) and (5) in (2)

$$E(s^2) = \frac{1}{n} \cdot \frac{n}{N} \sum_{i=1}^{N} Y_i^2 - \frac{1}{n(n-1)} \cdot \frac{n(n-1)}{N(N-1)} \sum_{i \neq j=1}^{N} Y_i Y_j$$

$$= \frac{1}{N} \sum_{i=1}^{N} Y_i^2 - \frac{1}{N(N-1)} \sum_{i \neq j=1}^{N} Y_i Y_j$$

$$= \frac{1}{N-1} \left[\sum_{i=1}^{N} Y_i^2 - N \bar{Y}_N^2 \right] \quad \text{Using .....(1)}$$

$$= S^2$$

$$E(s^2) = S^2$$

Therefore mean square is an unbiased estimate of the population mean square.

Theorem 3.7 (iii)

In srswor, the variance of the sample mean is given by

$$\text{Var}(\bar{y}_n) = \frac{S^2}{n} \cdot \frac{N-n}{N}$$

We have $\text{Var}(\bar{y}_n) = E(\bar{y}_n^2) - [E(\bar{y}_n)]^2$

$$= E(\bar{y}_n^{\,2}) - \bar{Y}_N^{\,2} \qquad [\because E(\bar{y}_n) = \bar{Y}_N] \ldots (1)$$

Now

$$E(\bar{y}_n^{\,2}) = E\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right]^2$$

$$= \frac{1}{n^2} E\left[\sum_{i=1}^{n} y_i^2 + \sum_{i\neq j=1}^{n} y_i y_j\right]$$

$$= \frac{1}{n^2}\left[E\left(\sum_{i=1}^{n} y_i^2\right) + E\left(\sum_{i\neq j=1}^{n} y_i y_j\right)\right] \qquad \ldots. (2)$$

We know

$$E\left(\sum_{i=1}^{n} y_i^2\right) = \frac{n}{N}\left(\sum_{i=1}^{N} Y_i^2\right)$$

But $\sum_{i=1}^{N}(Y_i - \bar{Y}_N)^2 = \sum_{i=1}^{N} Y_i^2 - N\bar{Y}_N^2$

Therefore $\sum_{i=1}^{N} Y_i^2 = \sum_{i=1}^{N}(Y_i - \bar{Y}_N)^2 + \mathrm{N}\,\bar{y}_N^2$

$= (\mathrm{N}\text{-}1)\mathrm{S}^2 + \mathrm{N}\,\bar{Y}_N^2$ \qquad But $\mathrm{S}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y}_N)^2$

$$(N-1)S^2 = \sum_{i=1}^{N}(Y_i - \bar{Y}_N)^2$$

$$\mathrm{E}\left(\sum_{i=1}^{N} Y_i^2\right) = \frac{n}{N}.\left[(N-1)S^2 + N\bar{Y}_N^2\right]$$

$$= \mathrm{n}\left[\frac{(n-1)}{N}S^2 + \bar{Y}_N^2\right] \qquad \ldots (3)$$

$$\text{Also } E\left(\sum_{i\neq j=1}^{n} y_i y_j\right) = \frac{n(n-1)}{N(N-1)} \sum_{i\neq j=1}^{N} Y_i Y_j$$

$$= \frac{n(n-1)}{N(N-1)} \left[\left(\sum_{i=1}^{N} Y_i\right)^2 - \sum_{i=1}^{N} Y_i^2\right]$$

$$= \frac{n(n-1)}{N(N-1)} \left[N\bar{Y}_N)^2 - [(N-1)S^2 + N\bar{Y}_N^2\right]$$

$$= \frac{n(n-1)}{N(N-1)} [\, N^2 \bar{y}_N^{-2} - (N-1)S^2 - N\bar{y}_N^{-2}]$$

$$= \frac{n(n-1)}{N(N-1)} [\, N(N-1)\, \bar{y}_N^{-2} - (N-1)S^2]$$

$$= n\,(n-1)\,[\bar{Y}_N^2 - \frac{1}{N} S^2] \qquad \ldots (4)$$

Substituting equations (3) and (4) in (2)

$$E(\bar{y}_n^{-2}) = \frac{1}{n^2} \left[n\left[\left(\frac{N-1}{N}\right)S^2 + \bar{Y}_N^2\right] + n(n-1)\left[\bar{Y}_N^2 - \frac{S^2}{N}\right]\right]$$

$$= \frac{n}{n^2}\left[(1-\frac{1}{N})S^2 + \bar{Y}_N^2 + (n-1)\bar{Y}_N^2 - (n-1)\frac{S^2}{N}\right]$$

$$= \frac{1}{n}\left[\left(\frac{(N-1)}{N}\right)S^2 + n\bar{Y}_N^2 - (n-1)\frac{S^2}{N}\right]$$

$$= \frac{(N-1)}{N}\left(\frac{S^2}{N}\right) + \bar{Y}_N^2 - \frac{(n-1)}{n}\cdot\frac{S^2}{N}$$

Substitute in equation (1)

The factor 1 – f is called finite population correction (f.p.c)

When $N \to \infty$ (Population size is large)

$$\text{f.p.c} = 1\text{-f} = 1 - \frac{n}{N} \to 1$$

$$\text{Therefore var } (\bar{y}_n) = \frac{S^2}{n}$$

Note 2:

The standard error of the sampling distribution of $(\bar{y}_n)$ is

$$\text{S.E } (\bar{y}_n) = \sqrt{\frac{N-n}{N}} \cdot \frac{S}{\sqrt{n}}$$

Note 3 :

When S is unknown replace S^2 by its unbiased s^2

$$\text{Therefore S.E } (\bar{y}_n) = \sqrt{\frac{N-n}{N}} \cdot \frac{s}{\sqrt{n}}$$

3.8 : SRSWOR VS SRSWR

Simple random sampling without replacement (SRSWOR)

Vs. Simples Random sampling with replacement (SRSWR)

Consider SRSWR from a population with variance σ^2

$$\text{Then var } (\bar{y}_n) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \text{var}(y_i)$$

The covariance terms vanish since in SRSWR all the draws are independent and consequently y_i's (i = 1,2,.....n) are

independently and identically distributed (i.i.d) with the same variance σ^2.

Therefore var $(\bar{y}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2$

$$= \frac{1}{n^2}.n\sigma^2 = \frac{\sigma^2}{n} \qquad \ldots.(1)$$

But $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y}_N)^2$

$$\Rightarrow N\sigma^2 = \sum_{i=1}^{N}(Y_i - \bar{Y}_N)^2$$

$N\sigma^2 = (N-1)S^2$ Q $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y}_N)^2$

Substitute in equation (1) $\sigma^2 = \frac{(N-1)}{N}S^2$

Therefore var $(\bar{y}_n) = \frac{(N-1)}{nN}S^2$

Note 5:

The variance of the sample is more in (SRSWR) as compared with its variance in the case of SRSWOR.

Therefore SRSWOR provides more efficient estimator of $(\bar{y}_n)$ relative to SRSWR.

3.9: SAMPLING ERRORS

Sampling error occurs when the part of the sample is selected from the population for enumeration of specified survey.

These errors can be rectified through complete enumeration. Normally in census, as complete enumeration is being done, so that sampling errors are absent in this case.

Sampling errors are primarily due to the following reasons.

i. **Faulty selection of the sample** :

Sometime defective sampling techniques are selected in the selection of sampling for example, in judgment sampling, the investigator deliberately selecting a representative sample to obtain certain results. This bias can be over come by selecting simple random sampling or by selecting a sample at random.

ii. **Substitution :**

Some times, the investigator may substitute some member in the enumeration of the certain survey. But the substitute may not be possessed the character of the original. This bias can be overcome until the original is included in the sample.

iii. Faulty demarcation of the sampling units

Sometimes, in fixing sampling units bias may occur. The area selection in a certain area for specified objective survey may be wrong. For example the area to be considered for agricultural research may be wrong. It depends on the investigator whether to include them in the sample or not.

iv. Error due to improper choice of the statistics for estimating population parameters.

Example :

If $x_1, x_2 \ldots\ldots x_n$ be a sample of n independent

observations then the sample variance $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 / n.$ as an estimate of the population variance σ^2 is biased where as the statistic $\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is an unbiased estimator of σ^2.

3.10 : NON SAMPLING ERRORS

Non sampling errors arise at the stages of observations, ascertainment and processing of data. Therefore these errors are present in both complete enumeration survey and the sample survey.

The data obtained in a complete census free from sampling errors. But still be subject to non-sampling errors. But data obtained in a sample survey will be subject to both sampling and non sampling errors.

Non sampling errors occurs at every stage of planning (or) execution of the sample survey.

Non-sampling errors arise from the following factors.

1.Faulty planning (or) definitions.

Planning of the survey involves in stating the objective of the survey. These set of objectives are then converted into (i) a set of definitions of the characteristics for which data ate to be collected and (ii) into a set of specifications for collection, processing and publications. Here non sampling errors may be due to.

a. Data specifications being inadequate and inconsistent with respect to the objective of the survey.

b. Errors due to location of units and actual measurement of characteristics, errors due to faculty questionnaire.

c. Lack of trained and qualified investigators.

2.Response errors :

These errors are introduced as a result of the responses furnished by the respondents any may be due to any of the following reasons.

i. Response errors may be accidental Eg. The respondent may misunderstand a particular question and accordingly furnish improper information.

ii. Prestige bias : Example : sometimes one may tell his about his education status, job, and salary to the interviewer to maintain his prestige.

iii. Self – interest : Sometimes respondents may hide their original data on their self interest. For (eg) The person may under estimate his salary, or production and over statement of this expenses.

Bias due to interviewer :

Sometimes interviewer may affect the accuracy of the response. The information obtained on suggestion from the interviewer is very likely to be influenced by the interviewer's beliefs and prejudices.

3. **Failure of memory** :

Sometimes respondent may fail to remember the events which will lead to the incomplete questionnaire. So that complete

enumeration of the survey is not possible.

4. **Non response bias** :

When the full information is not obtained from the sample units, non response bias occur. Sometimes the respondent may not be available in the house (or) may refuse to answer the questions (or) may be unable to furnish the details of the interviewer questions. In such cases, these bias cannot be avoided.

5. **Errors in coverage** :

Investigator has to decide the area limits for the specified survey proposed. Resources may be not be available, (or) the selected are may not be suitable (or) even the area limits are enough (or) too large to compute surveys. So the investigator must collect the details of the area limits from reliable sources.

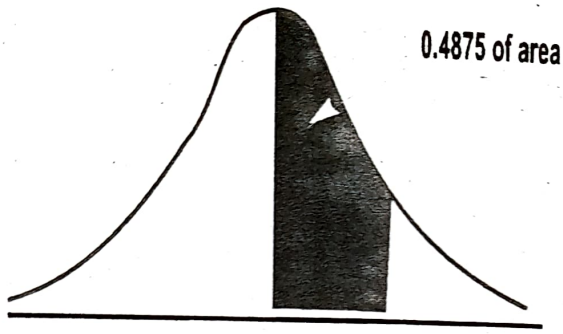
6. **Compiling errors** :

Once the survey has completely taken and when they are competed the errors may happen due to faulty entry, wrong calculations, missing of data. These errors can be rectified, through rechecking, the compiled data by the investigator.

7. **Publishing Errors :**

While the compiled survey data are sent to publishing houses for publishing work some contents may be misprinted (or) not printed, (or) some pages may be missed (or) overprinting, pages may be torn. These errors are identified in the proof and they are rectified before going into for final printing.

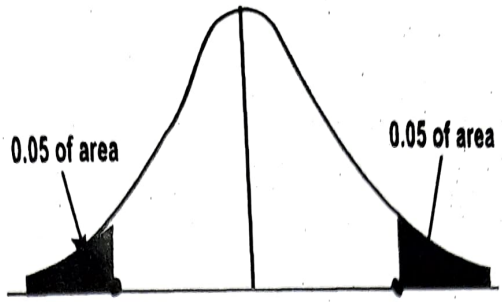
APPENDIX TABLES



Appendix Table 1

Areas under the Standard Normal Probability Distribution Between the Mean and Positive Values of z.

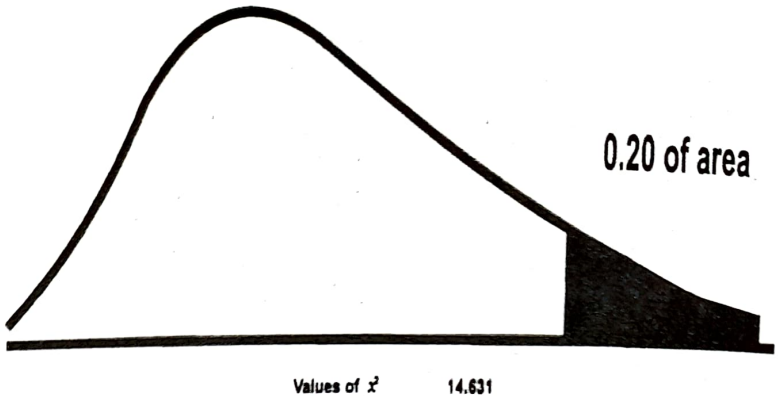
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4394	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4505	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4599	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4678	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4619	0.4656	0.4664	0.4671	0.4744	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4798	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4842	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4878	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4906	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4929	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4946	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4960	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4970	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4978	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4984	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4989	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990



Appendix Table 2

Areas in both tails combined for Student's t-Distribution

Degrees of Freedom	Area in Both Tails Combined			
	0.10	0.50	0.02	0.01
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771
28	1.708	2.048	2.467	2.763
29	1.671	2.045	2.462	2.756
30	1.699	2.042	2.457	2.750
40	1.697	2.021	2.423	2.704
60	1.684	2.000	2.390	2.660
120	1.658	1.980	2.358	2.617
Normal Distribution	1.645	1.960	2.326	2.576



Appendix Table 3

Area in the Right Tail of a Chi-sqyare (x^2) Distribution

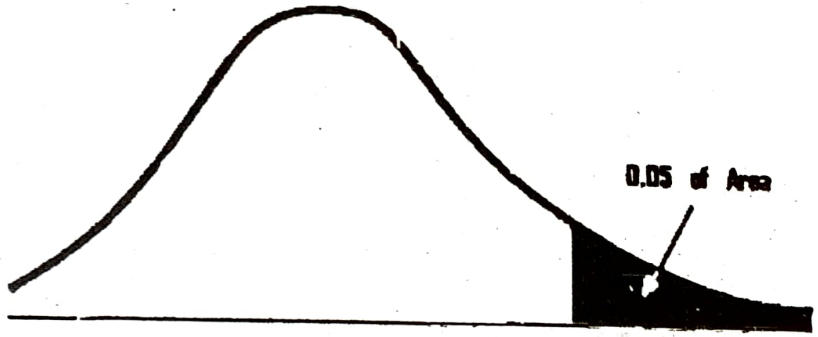
Degrees of Freedom	Area in Right Tail				
	0.99	0.975	0.95	0.90	0.800
1	0.00016	0.00098	0.00398	0.0158	0.0642
2	0.0201	0.0506	0.103	0.211	0.446
3	0.115	0.216	0.352	0.584	1.005
4	0.297	0.484	0.711	1.064	1.649
5	0.554	0.831	1.145	1.610	2.343
6	0.872	1.237	1.635	2.204	3.070
7	1.239	1.690	2.167	2.833	3.822
8	1.646	2.180	2.733	3.490	4.594
9	2.088	2.700	3.325	4.168	5.380
10	2.558	3.427	3.940	4.865	6.179
11	3.053	3.816	4.575	5.578	6.989
12	3.571	4.404	5.226	6.304	7.807
13	4.107	5.009	5.892	7.042	8.634
14	4.660	5.629	6.571	7.790	9.467
15	5.229	6.262	7.261	8.547	10.307
16	5.812	6.908	7.962	9.312	11.152
17	6.408	7.564	8.672	10.085	12.002
18	7.015	8.231	9.390	10.865	12.857
19	7.633	8.907	10.117	11.651	13.716
20	8.260	9.591	10.851	12.443	14.578
21	8.897	10.283	11.591	13.240	15.445
22	9.542	10.982	12.338	14.041	16.314
23	10.196	11.689	13.091	14.848	17.187
24	10.856	12.401	13.848	15.658	18.062
25	11.524	13.120	14.611	16.473	18.940
26	12.198	13.844	15.379	17.292	19.820
27	12.879	14.573	16.151	18.114	20.703
28	13.565	15.308	16.928	18.939	21.588
29	14.256	16.047	17.708	19.768	22.475
30	14.953	16.791	18.493	20.599	23.364

Appendix Table 3 (Continued...)

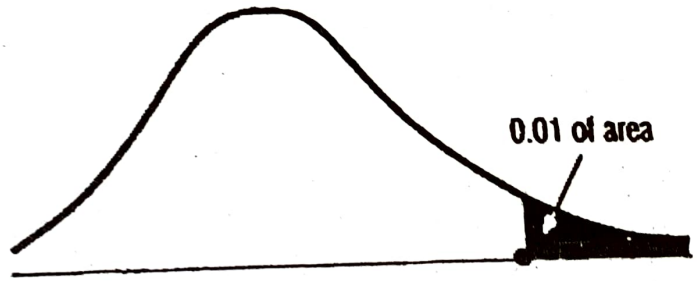
Area in Right Tail					Degrees of Freedom
0.20	0.10	0.05	0.025	0.01	
1.641	2.706	3.841	5.024	6.635	1
3.219	4.605	5.991	7.378	9.210	2
4.642	6.251	7.815	9.348	11.345	3
5.989	7.779	9.488	11.143	13.277	4
7.289	9.236	11.070	12.833	15.086	5
8.558	10.645	12.592	14.449	16.812	6
9.803	12.017	14.067	16.013	18.475	7
11.030	13.362	15.507	17.535	20.090	8
12.242	14.684	16.919	19.023	21.666	9
13.442	15.987	18.307	20.483	23.209	10
14.631	17.275	19.675	21.920	24.725	11
15.812	18.549	21.026	23.337	26.217	12
16.985	19.812	22.362	24.736	27.688	13
18.151	21.064	23.685	26.119	29.141	14
19.311	22.307	24.996	27.488	30.578	15
20.465	23.542	26.296	28.845	32.000	16
21.615	24.769	27.587	30.191	33.409	17
22.760	25.989	28.869	31.526	34.805	18
23.900	27.204	30.144	32.852	36.191	19
25.038	28.412	31.410	34.170	37.566	20
26.171	29.615	32.671	35.479	38.932	21
27.301	30.813	33.924	36.781	40.289	22
28.429	32.007	35.172	38.076	41.638	23
29.553	33.196	36.415	39.364	42.980	24
30.675	34.382	37.652	40.647	44.314	25
31.795	35.563	38.885	41.923	45.642	26
32.912	36.741	40.113	13.194	46.963	27
34.027	37.916	41.337	44.461	48.278	28
35.139	39.087	42.557	45.722	49.588	29
36.250	40.256	43.773	46.979	50.892	30

Appendix table 4 (a)

Values of F for F Distributions with 0.05 of the Area in the Righ



Degrees of Freedom for Denominator	Degrees of Freedom for Numerator																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	195
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	13.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	297	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00



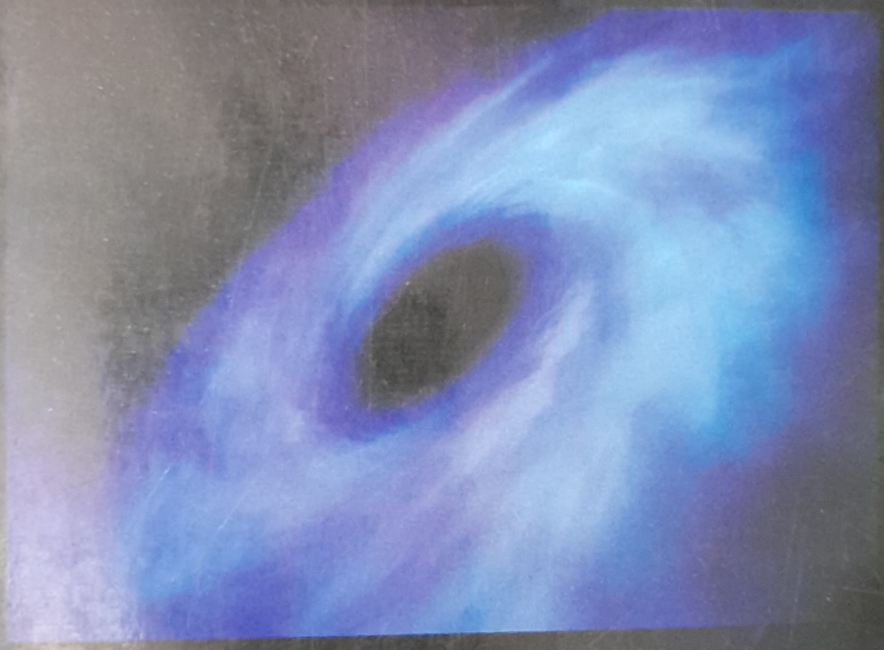
Appendix Table 4 (b)

Values of F for F Distributions with 0.01 of the Area in the Right Tail

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	30	40	60	120	∞
1	4,052	5,000	5,403	5,625	5,764	5,859	5,928	5,982	6,023	6,056	6,106	6,157	6,209	6,235	6,261	6,287	6,313	6,339	6,366											
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5											
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	26.1											
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	13.5											
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02											
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88											
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65											
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86											
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31											
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91											
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60											
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36											
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17											
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00											
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87											
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75											
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65											
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57											
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49											
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42											
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36											
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31											
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26											
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21											
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.53	2.45	2.36	2.27	2.17											
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01											
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80											
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60											
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38											
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00											

MATHEMATICAL APPLICATIONS OF TEST OF HYPOTHESIS AND SAMPLING TECHNIQUE

K.M. MANIKANDAN



Varun Publications
Ponnagar, Metturdam,
Salem - 634 601
E-mail : varunindu@yahoo.co.in

978-81-907472-3-3